Informe PEASIIS

Set de datos de investigación

Programa de Evaluación, Acreditación y Seguimiento de Institutos de Investigación Sanitaria

Subdirección General de Evaluación y Fomento de la Investigación

30 de enero de 2026

Versión 5.0

Grupo de trabajo "set de datos de investigación"

Miembros del grupo de trabajo

Dr. Luis García Ortiz. Director Científico del IIS IBSAL. Coordinador.

Dr. Francisco Tinahones Madueño.Director Científico del IIS IBIMA Plataforma BIONAND.

Dra. Michela Bertero. Directora de estrategia IIS IDIBAPS.

Dra. Pilar Rico Castro. Jefe de servicio y Coordinadora de Ciencia Abierta FECYT.

Dra. Remedios Melero Melero. Investigadora CSIC.

Dña. Concepción Campos Asensio. Bibliotecaria. Hospital Universitario de Getafe.

Dra. Cristina Lugones Sánchez. Responsable del área de gestión de datos del IBSAL.

Email de contacto:

Luis García Ortiz: direccioncientifica@ibsal.es

Francisco Tinahones: fjtinahones@uma.es

Michela Bertero: michela.bertero@idibaps.org

Pilar Rico: pilar.rico@fecyt.es

Remedios Melero: rmelero@iata.csic.es

Concepción Campos: ccampos@salud.madrid.org

Cristina Lugones: redcap@ibsal.es

Índice

1. Resumen ejecutivo

El Instituto de Salud Carlos III, dentro del marco del Programa de Evaluación, Acreditación y Seguimiento del IIS (PEASIIS), ha creado un grupo de trabajo con el fin de establecer un marco común para la recopilación, estructuración, almacenamiento, documentación y depósito de los sets de datos de investigación generados en proyectos dentro de los Institutos de Investigación Sanitaria (IIS) acreditados.
Este grupo de trabajo se constituyó el 12 de junio de 2025, donde se definieron las líneas principales a desarrollar en el documento. Tras ello, los miembros han trabajado en la elaboración y revisión de las sucesivas versiones del documento hasta la versión final del 10 de noviembre de 2025 (véase cronograma).
El presente documento propone un protocolo general para la gestión de datos de investigación siguiendo los principios FAIR (Findable, Accessible, Interoperable, Reusable), donde se establece un conjunto mínimo de información que deben incluir los sets de datos para incorporarlos a los repositorios, así como unos indicadores con el fin de evaluar las políticas de acceso abierto del ISCIII, así como la calidad de los sets de datos depositados.

Cronograma

12-6-2025	Constitución del grupo de trabajo
29-08-2025	Versión 1 del documento
24-09-2025	Versión 2 del documento
20-10-2025	Versión 3 del documento
07-11-2025	Versión 4 del documento
10-11-2025	Versión final del documento

2. Introducción

El Instituto de Salud Carlos III, dentro del marco del Programa de Evaluación, Acreditación y Seguimiento del IIS (PEASIIS), ha creado un grupo de trabajo con el fin de establecer un marco común para la recopilación, estructuración, almacenamiento y documentación de los sets de datos de investigación generados en proyectos dentro de los Institutos de Investigación Sanitaria (IIS) acreditados.
En el contexto del movimiento de Ciencia Abierta, la capacidad de compartir y reutilizar los conjuntos de datos se ha vuelto un imperativo. En este sentido, los principios FAIR (Findable, Accessible, Interoperable, Reusable) promueven que los datos sean localizables, accesibles bajo condiciones claras, compatibles con otros sistemas y reutilizables. Al depositar los sets de datos en repositorios especializados, los investigadores no solo aumentan la transparencia y el impacto de su trabajo, sino que también aceleran la innovación, permitiendo que otros exploren nuevas preguntas o verifiquen resultados existentes. Por lo tanto, el conjunto de datos de investigación no es un mero subproducto, sino un activo científico valioso por derecho propio. Sin embargo, son necesarias unas indicaciones que estandaricen el formato y la información que debe acompañar a estos set de datos para cumplir con estos principios. Por ello, el presente documento pretende establecer un estándar para la documentación y estructura de los

conjuntos de datos que se depositan en un repositorio. Su propósito es garantizar la coherencia, calidad y trazabilidad de la información, facilitando su comprensión, reutilización y mantenimiento a lo largo del tiempo. Al definir los elementos mínimos que cada set de datos debe incluir, se promueve una gestión de datos más eficiente, transparente y alineada con las mejores prácticas de organización y gobernanza de datos.

3. Objetivo

El objetivo general del documento es establecer un marco común para la recopilación, estructuración, almacenamiento y documentación de los sets de datos de investigación generados en proyectos dentro de los Institutos de Investigación Sanitaria (IIS) acreditados.

Este procedimiento busca garantizar la calidad, interoperabilidad, trazabilidad y reutilización de los datos, facilitando su integración en repositorios institucionales o temáticos en acceso abierto, siempre que sea posible. Asimismo, promueve el cumplimiento de directrices éticas y normativas, así como de los principios FAIR (Findable, Accessible, Interoperable, Reusable). También contribuirá a la transparencia, eficiencia y colaboración en la investigación científica, favoreciendo la reproducibilidad y el impacto de los resultados.

De forma específica se definen los siguientes objetivos:

O1. Diseñar un procedimiento o protocolo general común para el almacenamiento y gestión de los datos de investigación en todos los IIS, incluyendo:

- Fases de recogida: fuente de datos, periodicidad, almacenamiento, etc.
- Herramientas recomendadas para la gestión técnica de los datos.
- Procedimiento/s para el depósito en abierto, siempre que sea posible, en repositorios nacionales o internacionales.

O2. A partir del marco común desarrollado establecer criterios para la estandarización y definición de un/os indicador/es de calidad de los conjuntos de datos de investigación susceptible de ser depositados en repositorios abiertos, aplicable de forma homogénea en todos los IIS acreditados y a otros centros del Sistema Nacional de Salud (SNS).

4. Alcance

El presente procedimiento es de aplicación para todos los conjuntos de datos generados, recolectados, transformados o utilizados por los IIS en el marco de actividades de investigación científica, técnica o de desarrollo experimental, independientemente de su área o de la metodología empleada para su obtención. Abarca todas las fases del ciclo de vida de los datos: planificación, recolección, validación, procesamiento, análisis, documentación, curación, preservación, publicación, depósito y reutilización.

Se incluyen los datos primarios y secundarios generados por investigadores, personal técnico y colaboradores que participen en proyectos o unidades de investigación con vinculación a los IIS. También son objeto de este procedimiento los datos derivados de proyectos en colaboración nacional o internacional donde la institución figure como entidad coordinadora, socia o participante. Se excluyen expresamente los datos generados con fines exclusivamente administrativos o de gestión interna no relacionados con investigación.

Este protocolo se alinea con el marco normativo vigente en materia de ciencia abierta, gestión de datos y protección de la información, incluyendo:

- Ley 14/2011 de la Ciencia, la Tecnología y la Innovación (modificada por la Ley 17/2022), que promueve el acceso abierto a resultados de investigación públicos
- Reglamento (UE) 2016/679 (RGPD) y Ley Orgánica 3/2018 (LOPDGDD), en materia de protección de datos personales.
- Decreto 24/2021 de Transposición de 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.
- Reglamento UE 2021/695 (Horizonte Europa), que establece la obligatoriedad del generar un plan de gestión de datos (PGD) y del acceso a los datos tan abierto como sea posible y tan cerrado como sea necesario.
- Anotated Grant Agreement for EU Funded Programmes 2021-27 (https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/common/guidance/aga_en.pdf)
- Ley Orgánica 2/2023, del Sistema Universitario, donde se especifican acciones y procedimientos para el Fomento de la Ciencia Abierta y Ciencia Ciudadana.
- Estrategia Nacional de Ciencia Abierta 2023–2027 (ENCA), que instaura la gestión de datos de investigación bajo principios FAIR.
- Plan Estatal de I+D+I, que requiere la presentación de Planes de Gestión de Datos (PGD).

A nivel institucional, el ISCIII ha desarrollado:

- Guías técnicas para la elaboración de PGD, incluyendo herramientas para los IIS
- El repositorio Repisalud (https://repisalud.isciii.es/home), que implementa políticas de ciencia abierta para publicación de datos y metadatos siguiendo los principios FAIR (https://fairsharing.org/)
- Un marco específico para datos sensibles y sanitarios, con protocolos de anonimización, procedimientos de acceso, evaluación de impacto y licencias compatibles

La estandarización/normalización de los conjuntos de datos responde a la necesidad de garantizar su calidad, trazabilidad, interoperabilidad y reutilización, en línea con estándares internacionales y requerimientos de organismos financiadores. La dispersión en formatos, estructuras y esquemas de metadatos dificulta la integración, el acceso y la reproducibilidad de los resultados científicos, y compromete la visibilidad, impacto y evaluación de la producción institucional. Este procedimiento establece instrucciones

técnicas para la comunidad investigadora y los IIS con el fin de reducir dicha heterogeneidad, facilitar la implementación de políticas de ciencia abierta y asegurar una gestión eficiente, ética, responsable y sostenible de los datos de investigación. Asimismo, permite una evaluación más exhaustiva y objetiva del cumplimiento de los requerimientos de ciencia abierta.

5. Definiciones

- **Datos de investigación**: Conjunto de registros, observaciones, mediciones, resultados experimentales, transcripciones, códigos, imágenes u otros elementos digitales o analógicos que sirven como base para validar, reproducir y derivar conclusiones científicas. Incluyen tanto datos primarios como secundarios.
- **Set de datos (dataset en inglés)**: Colección de información organizada en un formato estructurado que permite su comprensión, procesamiento y análisis.
- **Metadatos**: Información descriptiva (detalles de los datos), estructural (cómo se organiza el set de datos), técnica (tipo de dato, y forma de obtención), administrativa (acceso y gestión del set de datos) y contextual (información relevante adicional) que documenta los datos de investigación, facilitando su identificación, interpretación, preservación y reutilización.
- **Plan de Gestión de Datos (PGD o DMP en inglés)**: Documento formal que describe cómo se gestionarán los datos durante y después del proyecto de investigación, incluyendo aspectos de almacenamiento, preservación, acceso y protección legal o ética.
- **Principios FAIR**: Conjunto de directrices internacionales que buscan que los datos de investigación sean Localizables (Findable), Accesibles (Accessible), Interoperables (Interoperables) y Reutilizables (Reusable). Requieren el almacenamiento en repositorios de confianza, el uso de identificadores persistentes, estándares de interoperabilidad abiertos, la aplicación de licencias de acceso y reutilización claras y metadatos detallados.
- **Datos sensibles**: Datos que, por su naturaleza, pueden comprometer la privacidad, la seguridad o los derechos de personas físicas, colectivos o entidades. En investigación sanitaria, esto incluye datos clínicos, genéticos o de salud, los cuales están protegidos por el RGPD y la LOPDGDD.
- **Anonimización**: Proceso técnico consistente en eliminar o transformar de forma irreversible los datos personales para que no sea posible identificar, directa ni indirectamente, a la persona a la que se refieren. Es requisito obligatorio para la reutilización pública de datos sensibles según la normativa europea.
- **Pseudonimización**: Proceso técnico consistente en el tratamiento de los datos personales para que no puedan atribuirse a un sujeto concreto sin utilizar información adicional. Dicha información adicional debe almacenarse por separado y estar sujeta a medidas técnicas y organizativas destinadas a garantizar la identificación de la persona física.
- **Repositorio**: Infraestructura digital, gestionada por una institución o comunidad científica, destinada al almacenamiento, preservación y acceso abierto a los

resultados de investigación, incluidos los sets de datos y sus metadatos. Estos repositorios pueden ser institucionales, temáticos o multidisciplinares.

- **Licencia de uso**: Instrumento legal mediante el cual, el titular de los derechos de propiedad intelectual permite a terceras personas realizar determinados usos de los datos. En el caso de set de datos, se recomienda utilizar licencias abiertas del tipo Creative Commons o las Open data Commons.

6. Protocolo general de los sets de datos

1. Responsabilidades en materia de datos

En el contexto de la ciencia abierta y la gestión responsable de los datos de investigación, es necesaria una distribución de responsabilidades entre las instituciones de investigación y los equipos científicos. Con ello se garantiza que los datos generados durante el proceso investigador cumplan con los requisitos legales y éticos, estén estructurados, documentados y preservados de forma que faciliten su acceso, interoperabilidad y su posible reutilización.

Instituto de investigación sanitaria

El Instituto de investigación sanitaria, como entidad responsable del entorno institucional de investigación, debe establecer un marco normativo y técnico que garantice la correcta gestión de los datos de investigación, especialmente aquellos que contienen información sensible o de carácter personal. Entre sus responsabilidades se incluyen el desarrollo de políticas de protección de datos alineadas con la normativa vigente, así como con estándares de interoperabilidad y preservación digital. La institución es responsable de proporcionar infraestructuras seguras y sostenibles para el almacenamiento y procesamiento de datos. Asimismo, debe ofrecer formación especializada y recursos técnicos que faciliten la implementación adecuada de los planes de gestión de datos (PGD) en cada proyecto, así como mecanismos de auditoría y control para asegurar la trazabilidad y la integridad de los datos durante todo el ciclo de vida del proyecto.

Equipo investigador

El investigador/a principal asume la responsabilidad directa sobre la generación, tratamiento y documentación de los datos dentro del proyecto, garantizando su alineación con los principios FAIR desde el diseño del estudio hasta la fase de publicación y preservación. Esto implica elaborar un Plan de Gestión de Datos que contemple no solo los aspectos éticos y legales del tratamiento, sino también su estructuración semántica, su interoperabilidad con estándares internacionales y su posterior disponibilidad en repositorios adecuados. El equipo investigador debe asegurar que los datos sean correctamente descritos mediante metadatos normalizados, aplicando técnicas de anonimización cuando corresponda, y facilitando su reutilización futura. También le corresponde la notificación de incidentes relacionados con la

seguridad de los datos y la coordinación con la institución para garantizar su correcta conservación o eliminación al finalizar el proyecto.

2. Fases de recogida de datos

a. Fuentes de datos

- Fuentes internas: datos experimentales, clínicos, encuestas, entrevistas, registros electrónicos, biobancos.
- Fuentes externas: bases de datos públicas, sistemas de información del SNS, cohortes internacionales, literatura científica.
- Clasificación por tipo de datos: estructurados (bases de datos, tablas), no estructurados (imágenes, audio, texto), datos sensibles/personales.

b. Periodicidad

- Continua: sistemas de monitorización, sensores, EHR.
- Periódica: mensual, trimestral, según fases descritas en el proyecto.
- Puntual: recogida única (ej. encuestas o muestras biológicas).

c. Estandarización en la recogida

- Uso de formularios electrónicos normalizados (REDCap, OpenClinica).
- Validación automática de campos y control de calidad.
- Asignación de códigos únicos y pseudonimización desde el origen.

d. Almacenamiento inicial

- Almacenamiento seguro en servidores que cumplan con estándares de seguridad de la información, protección de datos personales y preservación digital.
- Cifrado de datos sensibles en tránsito y en reposo.
- Registro de logs de acceso, auditoría de modificaciones y backups programados.

3. Herramientas recomendadas para la gestión técnica de los datos

Área funcional	**Herramientas sugeridas**	**Observaciones**
Recolección/entrada	REDCap, LimeSurvey, OpenClinica	Cumplen normativas GDPR/LOPDGDD.
Gestión de bases de datos	PostgreSQL, MySQL, MongoDB	Adaptables a distintos tipos de datos y escalables.
Anonimización/pseudonimización	ARX, Amnesia, SDCMicro	Herramientas validadas para datos clínicos y biomédicos.

Área funcional	Herramientas sugeridas	Observaciones
Control de versiones y trazabilidad	Git, DVC (Data Version Control)	Recomendado para proyectos colaborativos.
Estandarización de metadatos	Dataverse, CKAN, Invenio	Softwares para repositorios de datos que soportan esquemas como DataCite, Dublin Core, OpenAIRE, DCAT.
Documentación del PGD	DMPTool, Argos (OpenAIRE), OpenDMP, DMPonline	Sugeridas por agencias financiadoras (ej. Comisión Europea).

4. Procedimiento para el depósito en repositorios abiertos

a. Selección del repositorio adecuado

- Buscador de repositorios de datos: https://www.re3data.org/
- Repositorios institucionales. Ejemplos: Repisalud, DIGITAL.CSIC.
- Repositorios temáticos. Ejemplos: EGA (datos genómicos), Biomodels.
- Repositorios multidisciplinares. Ejemplos: Figshare, Zenodo (véase Anexo I).

b. Requisitos previos al depósito

- Datos mínimos para la reproducibilidad del estudio, anonimizados o con consentimiento para su difusión.
- Inclusión de metadatos normalizados (Dublin Core, OpenAIRE o DataCite).
- Licencia clara y preferiblemente abierta.
- Identificador persistente generado por el repositorio o asignado externamente (DOI, Handle, ARK, URN, etc.).
- Revisión interna por comité ético o unidad de datos (según caso).

c. Proceso de depósito

Preparación del set de datos final:

- Estructura clara, se recomienda formato abierto (CSV, JSON, XML, etc.).
- Denominación correcta de los archivos a subir:
  - El nombre de los archivos debe ser conciso, informativo y único. Debe incluir: fecha del fichero, nombre del proyecto, test (si hay varios), tipo de datos o de análisis, identificador de la persona responsable de su creación, versión del fichero y extensión.
  - Evitar espacios en blanco y caracteres especiales, ya que pueden producir errores. Se pueden usar patrones como los siguientes:
    CamelCase: 20250108HGPTest1PMartinezV3. xlsx

Snake_case: 20250108_HGP__test1_PMartinez_v3. xlsx

- Documentación técnica asociada: Código de variables, metodología, scripts de análisis (si aplica) y archivo README.

Carga en repositorio:

- Subida manual o automatizada (según herramienta).
- Revisión y validación de metadatos.

Publicación y citación:

- Generación de cita recomendada.
- Notificación al equipo del proyecto y al Instituto de investigación en la memoria científica del proyecto

d. Mantenimiento post-depósito

- Actualización en caso de nueva versión del set de datos.
- Registro del impacto y reutilización (altmetrics, citas).
- Comprobar los mecanismos para reportar errores o inconsistencias del repositorio seleccionado.

7. Estructura y estandarización del set de datos

Criterios de estandarización

Con base en el protocolo común definido y conforme a los principios FAIR, se establecen los siguientes **criterios mínimos obligatorios** que debe cumplir un set de datos para ser considerado **estandarizado y apto para el depósito en repositorios abiertos**:

a. Criterios técnicos

- **Formato abierto**: se recomienda que el set de datos esté en formatos no propietarios (ej. .csv, . json, .tsv, .xml, .rdf, .txt, .tiff). (véase Anexo II)
- **Estructura clara**: los datos deben estar organizados en una estructura jerárquica o de ficheros, con relaciones documentadas.
- **Consistencia y validación**: se debe garantizar la ausencia de valores inconsistentes, errores de codificación o duplicidades.
- **Control de versiones**: se debe conservar trazabilidad de cambios.

b. Criterios semánticos

- **Diccionario de variables**: definición completa de cada campo, unidad de medida, codificación y código de variables. (véase Anexo III)
- **Uso de vocabularios controlados**: adopción de terminologías estandarizadas (ej. ICD-10, MeSH). (véase Anexo IV)
- **Estructura de metadatos**: mínimo requerido en Dublin Core, OpenAIRE o DataCite. (véase Anexo V)

c. Criterios legales y éticos

1. **Anonimización o pseudonimización** (véase Anexo VI)
2. **Consentimiento informado,** en caso necesario**.**
3. **Licencia de uso clara**: preferentemente aquellas que maximicen su reutilización. Herramientas como la de Creative Commons puede ayudarnos a seleccionar la más indicada dependiendo de las características de los datos.
4. **Evaluación ética**: Aprobación o informe del comité de ética del proyecto, especialmente cuando se traten de datos sensibles.

d. Criterios de documentación

- **Plan de Gestión de Datos (PGD)** completado, actualizado y accesible.
- **Archivo README**: documento adjunto con descripción del proceso de recogida, estructura del set de datos, componentes del equipo de trabajo, responsable de la gestión de datos y herramientas utilizadas para su tratamiento, entre otra información (véase anexo VII).

8. Indicadores para la evaluación del set de datos

La evaluación de estos indicadores se propone realizarla una vez se presente la memoria científica del proyecto, que deberá incluir información sobre los sets de datos depositados, con la siguiente estructura:

Campo	**Descripción**
Código y proyecto asociado	
IP del proyecto	
ORCID del IP	
Plan de gestión de datos (PGD) publicado	Sí / No
En caso afirmativo, URL o identificador del PGD	
Nombre del set de datos depositado	
Repositorio utilizado	
Identificador persistente asignado al set de datos	
Fecha de creación	dd/mm/yyyy
Tipo de datos	Imagen, texto, tabular, genómico, etc.
Formato de archivo/s	CSV, JSON, TIFF, etc.
Incluido archivo README y/o diccionario de variables	Sí / No
Incluidos metadatos mínimos *(Ver anexo V)*	Sí / No
Derechos de acceso	Público/Restringido/Embargo/Otros

Si hubiera más de un depósito de datos del proyecto, con identificador persistente diferente, se recomienda cumplimentar una ficha por cada uno de ellos. Los metadatos mínimos se solicitan para el registro en la mayoría de los repositorios de datos (p.ej. Dryad, Zotero, Figshare, OSF).

Indicadores

Indicador	Descripción	Valor esperado	Observaciones
A1. Implementación	% de proyectos financiados por ISCIII con al menos un set de datos publicado en un repositorio de datos	≥70%	Evaluación en la memoria científica presentada tras finalización del proyecto
A2. Existencia de un PGD	% de proyectos financiados por ISCIII con un PGD depositado en repositorios de acceso abierto	≥ 70%	Evaluación en la memoria científica presentada tras finalización del proyecto
B1. Formato de archivo abierto	% de sets de datos en formatos abiertos reutilizables (CSV, JSON, etc.)	≥ 70%	Incluye los formatos recomendados en el Anexo II.
B2. Presencia de metadatos mínimos	% de sets de datos con metadatos, como mínimo, los requeridos por el repositorio para el depósito.	≥ 90%	Verificable automáticamente
B3. Licencia de uso definida	% de sets de datos con licencia explícita	≥ 90%	Preferencia por licencias abiertas
B4. Documentación del set de datos	% de sets de datos con código de variables y/o documentación técnica	≥ 85%	Adjunta al depósito
B5. Anonimización completa	% de sets de datos con riesgo mínimo de reidentificación	100%	Requiere validación por parte del investigador previa al depósito
B6. Identificador persistente asignado	% de sets de datos con identificador persistente asignado	≥ 95%	Evaluado en repositorio destino

Aclaraciones:

*Los **indicadores A** evalúan el cumplimiento de la política de datos, por lo que el denominador del indicador son todos los proyectos finalizados con financiación del ISCIII del IIS que hayan presentado la memoria científica en el momento de la evaluación.*

*Los **indicadores B** evalúan la calidad de los sets de datos depositados en los repositorios de datos. Por ello, el denominador es el número de sets de datos que provengan de proyectos finalizados con financiación del ISCIII, y que hayan presentado la memoria científica en el momento de la evaluación.*

Aplicabilidad y evaluación transversal

- Estos indicadores, todos o algunos de ellos, se deberían integrar gradualmente en el sistema de **calidad y seguimiento de datos** de los IIS. También sería necesario capacitar de forma paralela a los IIS e investigadores para la recolección de la información y cálculo de indicadores.
- La evaluación de estos indicadores se puede automatizar parcialmente mediante herramientas como **FAIR Evaluation Services**, **RO-Crate** así como a través de los validadores de metadatos que implementan algunos repositorios.

ANEXO I. Zenodo como ejemplo de repositorio multidisciplinar

Zenodo es repositorio de acceso abierto, multidisciplinario, financiado por la Comisión Europea a través de la iniciativa OpenAIRE, y operado por el CERN (Organización Europea para la Investigación Nuclear), que permite depositar y preservar materiales resultados de proyectos de investigación, sin limitar su uso por procedencia de los usuarios. Alberga un amplio espectro de tipología documental (datos, publicaciones, software, informes, tesis, PGD, etc.) en su mayoría en abierto, pero también permite la opción de mantenerlos en cerrado siempre que sea necesario, y tan solo exponer los metadatos.

Además, cada depósito en Zenodo recibe un DOI (Digital Object Identifier) único y persistente. Esto asegura que la investigación depositada sea fácilmente citable y se mantenga accesible a largo plazo, incluso si la ubicación original del archivo cambia.

Zenodo permite el depósito de una gran variedad de archivos, incluyendo:

- Conjuntos de datos (sets de datos): Desde datos experimentales hasta encuestas y simulaciones.
- Publicaciones científicas: Artículos, preprints, informes técnicos y tesis.
- Software de investigación: Código fuente, scripts y aplicaciones.
- Materiales de presentación: Presentaciones en congresos y conferencias.
- Informes: Documentos internos, resúmenes de proyectos.
- Imágenes y videos: Material audiovisual asociada a los proyectos de investigación.

Además, ofrece integración con GitHub, permitiendo conservar archivos, asignándoles un DOI y asegurando su preservación. También soporta el uso de metadatos estandarizados, lo que mejora la identificación de los depósitos a través de recolectores y agregadores de contenido.

Las comunidades de datos en Zenodo

Permite crear comunidades científicas, lo que facilita el almacenamiento conjunto de datos e información de grupos de investigadores, bien por socios de proyecto, grupo de investigación o centros. En el caso de los IIS permite crear una comunidad especifica del Instituto en la que recopilar todos los datos y documentos que los investigadores depositan en el repositorio, siendo así fácilmente accesibles e identificables para la gestión de estos por los IIS.

Almacenamiento y seguridad

Proporciona un almacenamiento seguro y fiable para los datos de investigación. Cada usuario puede depositar hasta 50 GB por set de datos (aunque es posible solicitar más espacio para necesidades especiales). Los datos se almacenan en la infraestructura de CERN, garantizando altos estándares de seguridad y redundancia.

ANEXO II: Formatos de archivos

Tipo	Formatos recomendados	Formatos no recomendados, pero habitualmente aceptados
Documentos de texto	PDF/A (.pdf) ODT (.odt)	Microsoft Word (.doc) Office Open XML (.docx) Rich Text File (.rtf) PDF diferente a PDF/A (.pdf)
Texto plano	Unicode text (.txt)	No-Unicode text (.txt)
Archivos comprimidos	ZIP (.zip) 7-Zip (.7z) GNU Tar (tar.gz/.tgz) GNU Gzip (.gz)	WinRAR (.rar)
Lenguaje de marcado	XML (.xml) HTML (.html) Related files: .css, .xslt, .js, .es Markdown (.md)	SGML (.sgml)
Lenguajes de programación	NetCDF Text-Fabric Python R	MATLAB
Hojas de cálculo	ODS (.ods) CSV (.csv)	Microsoft Excel (.xls) Office Open XML Workbook (.xlsx) PDF/A (.pdf)
Bases de datos	SQL (.sql) SIARD (.siard) CSV (.csv)	Microsoft Access (.mdb, .accdb) dBase (.dbf) HDF5 (.hdf5, .he5, .h5)
Datos estadísticos	General: .dat/ .txt JASP (.csv/.html) R (.RData/.Rds)	SPSS (.sav) STATA (.dta/.DO) SAS (.7dat; .sd2; .tpt) JASP (.jasp)
Imágenes (mapa de bits)	JPEG (.jpg, .jpeg) TIFF (.tif, .tiff) PNG (.png) JPEG 2000 (.jp2) DICOM (.dcm)	
Imágenes vectoriales	SVG (.svg)	Adobe Illustrator (.ai) EPS (.eps) WMF/EMF (.wmf, .emf) CDR (.cdr)
Video	MXF (.mxf) Matroska (.mkv)	MPEG-4 (.mp4, .m4a, .m4v, …) MPEG-2 (.mpg, .mpeg, .m2v, .mpg2, …) AVI (.avi), QuickTime (.mov, .qt)
Audio	BWF (.bwf) MXF (.mxf) Matroska (.mka)	WAVE (.wav) MP3 (.mp3) AAC (.aac, .m4a)

	FLAC (.flac) OPUS	AIFF (.aif, .aiff) OGG (.ogg)
Diseño asistido por ordenador (CAD)	AutoCAD DXF version R12 (ASCII) (.dxf) SVG (.svg)	AutoCAD DXF, version different than R12 (ASCII) (.dxf) DWG (.dwg) DGN (.dgn)
Sistemas de información geográfica (GIS)	GML (.gml) MIF/MID (.mif/.mid) GeoJSON (.json) GeoPackage (.gpkk)	Esri Shapefiles (.shp & related files) MapInfo (.tab & related files) KML (.kml, .kmz) Esri Geodatabase (.gdb) Project files / Workspaces (.mxd, .wor, .qgs)
Imágenes georeferenciadas	GeoTIFF (.tif, .tiff)	TIFF World File (.tfw & .tif, possibly with additional files) JPEG World File (.jgw & .jpg, possibly with additional files) ERDAS IMAGINE File Format (.img)
GIS Raster	ASCII GRID (.asc, .txt)	Esri GRID (.grd & related files) Surfer Grid (.grd; .srf) ERDAS IMAGINE File Format (.img)
3D	WaveFront Object (.obj) Polygon file format (.ply) X3D (.x3d) glTF 2.0 (.gltf; .glb) COLLADA (.dae) LASer (.las, .laz) IFC (.ifc)	Autodesk FBX (.fbx) Blender (.blend) glTF 1.0 (.gltf; .glb) 3D PDF (.pdf) Google Draco (.drc) Artec (.a3d) Agisoft Metashape (.psx & .psz) STL (.stl) VRML (.wrl, .wrz, .vrml)
RDF	RDF/XML Trig (.trig) Turtle (.ttl) NTriples (.nt) JSON-LD	
Computer Assisted Qualitative Data Analysis (CAQDAS)	REFI-QDA (Qualitative Data Analysis)	
Datos de proteómica	MSF (.msf)	
Estándar de citometría de flujo	FCS (.fcs)	

Tabla elaborada a partir de la tabla de formatos de Data Archiving and Networked Services (DANS) y de la Gestión de dades de recerca.

ANEXO III: Diccionario de variables. Ejemplos

Variable: ocupación	
Código	**Etiqueta**
1	Estudiante
2	Trabajador por cuenta ajena
3	Trabajador por cuenta propia
4	Desempleado/a
5	Jubilación

Variable: date_infecc	
Formato	**Descripción**
AAAA/MM/DD	Año, mes y día en que la persona presentó el inicio de síntomas o recibió el diagnóstico de la infección.

Variable: imc	
Formato	**Descripción**
Número decimal con dos cifras (ej. 23,45)	Valor que relaciona el peso y la estatura. Se calcula como peso (kg) dividido entre la estatura (m) elevada al cuadrado.

ANEXO IV: Vocabularios controlados por disciplina

Disciplina	Nombre del vocabulario controlado	Ejemplo de término aceptado	Ejemplo de URL aceptado
Multidisciplinar	Library of Congress Subject Headings	Psychology	http://id.loc.gov/authorities/subjects/sh85108459
Multidisciplinar	DBpedia	Artificial intelligence	https://dbpedia.org/page/Category:Artificial_intelligence
Alimentación	FoodOn Food Ontology	maize flour	http://purl.obolibrary.org/obo/FOODON_03540073
Artes y Humanidades	Art and Architecture Thesaurus	aqueducts	http://vocab.getty.edu/page/aat/300006165
Artes y Humanidades	Tesauro de Arte & Arquitectura	acueducto	https://www.aatespanol.cl/terminos/300006165
Artes y Humanidades	Getty Thesaurus of Geographic Names	Iberian Peninsula	http://vocab.getty.edu/page/tgn/7016676
Artes y Humanidades	Tesauros-Diccionarios del patrimonio cultural de España	Ánfora	http://tesauros.mecd.es/tesauros/ceramica/1010577
Artes y Humanidades	Pleiades	Tarraconensis	https://pleiades.stoa.org/places/981551
Artes y Humanidades	PeriodO	Roman Imperial	http://n2t.net/ark:/99152/p0qhb66dt5q
Ciencias Agrarias	AGROVOC Multilingual Thesaurus	feed additives	http://aims.fao.org/aos/agrovoc/c_2827
Ciencias de la Vida	Medical Subject Headings (MeSH)	Biotechnology	https://www.ncbi.nlm.nih.gov/mesh/68001709
Ciencias de la Vida	International Classification of Diseases (ICD)	Pneumonia	http://id.who.int/icd/entity/142052508
Ciencias de la Vida	Gene Ontology	Lysosome	https://amigo.geneontology.org/amigo/term/GO:0005764
Ciencias de la Vida	Descriptors en ciències de la salut (bvsalud)	Capacidad de Absorbancia de Radicales de Oxígeno	https://decs.bvsalud.org/es/ths/resource/?id=56697&filter=ths_termall&q=Capacidad%20de%20Absorbancia%20de%20Radicales%20de%20Ox%C3%ADgeno
Ciencias de la Vida	The thesaurus on Prehistory	Anthracology	https://www.archeoindex.org/776

Tabla elaborada a partir de la tabla de vocabularios controlados de la Gestión de dades de recerca.

ANEXO V: Estructura de los metadatos según Dublin Core

Elemento	Descripción	Ejemplo
Title	Título del recurso	"Datos de prevalencia de diabetes en adultos - España 2023"
Creator	Autor o entidad responsable de la creación	"Instituto de Investigación Sanitaria X"
Subject	Tema o palabras clave (preferentemente usando vocabularios controlados)	"Diabetes tipo 2; epidemiología; salud pública"
Description	Descripción breve del contenido del set de datos	"Este conjunto contiene datos anonimizados de prevalencia de diabetes recogidos entre enero y diciembre de 2023 en centros de atención primaria."
Publisher	Entidad responsable de la publicación del recurso	"IIS"
Contributor	Otros colaboradores relevantes	"Departamento de Estadística, Universidad Y"
Date	Fecha asociada al recurso (publicación, creación, etc.)	"2024-02-15"
Type	Tipo de recurso según vocabulario controlado (DCMI Type Vocabulary)	"Set de datos"
Format	Formato del archivo, preferiblemente MIME type	"text/csv"
Identifier	Identificador único del recurso (preferentemente DOI, Handle, URI)	"https://doi.org/10.5281/zenodo.9999999"
Source	Recurso del que deriva, si aplica	"Encuesta Nacional de Salud 2020"
Language	Idioma del contenido principal (ISO 639-1)	"es"
Relation	Recursos relacionados (otros sets de datos, artículos, etc.)	"Relacionado con: https://doi.org/10.1016/j.pubmed.2024.00123"
Coverage	Alcance y/o espacio temporal	"España, 2023"
Rights	Condiciones de uso y licencia	"CC BY 4.0"

ANEXO VI. Anonimización de los datos

El set de datos no puede incluir identificadores directos, es decir, información que es suficiente por sí misma para identificar al paciente (nombre, dirección, código postal, número de teléfono, fotografías o videos). Se admite la inclusión de identificadores indirectos, aquellos que combinados podrían llevar a la identificación de la persona, pero los mínimos posibles, algunos autores recomiendan máximo 3. Algunos ejemplos serían afiliación, ocupación, sexo o medidas antropométricas.

La anonimización generalmente implica la eliminación de identificadores directos y la modificación de identificadores indirectos. La seudonimización, o codificación, se refiere a la sustitución de datos de identificación por valores ficticios, sin relación con los originales. Los seudónimos pueden ser irreversibles cuando los valores originales se eliminan correctamente y la seudonimización se realiza de una manera no repetible.

Cualquier variable que no sea necesaria en el set de datos debe eliminarse para minimizar los datos. La adecuada selección de las variables se realizaría según la figura:

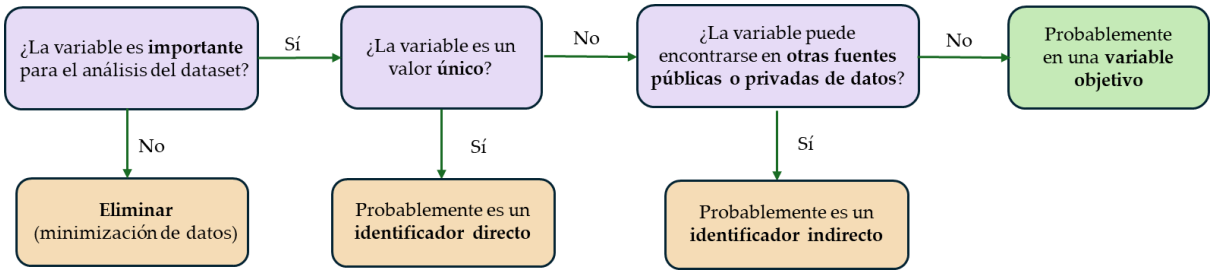

Figura elaborada a partir de la guía básica de anonimización de la AEPD

Existen muchas técnicas de anonimización que pueden utilizarse para modificar las variables del set de datos:

Técnica	Descripción	Ejemplos
Agregación	Transformar los datos en rangos	Utilizar rangos de edad en lugar de la fecha de nacimiento o la edad exacta (p.ej. 10-20)
Combinación	Fusionar dos variables o más creando una nueva variable resumen	Cambiar las fechas exactas por intervalos de tiempo entre eventos (p.ej. la duración del tiempo que pasó en una encuesta)

Tratamiento de outliers	Agrupar en rangos amplios una variable continua para evitar la identificación de valores atípicos	Una persona de 72 años se incluye en «personas de 70 años» o «personas mayores».
Generalización	Ajustar la precisión de los datos	Puesto profesional específico → ocupación o área de especialización Ciudad → región
Limitación del número de identificadores indirectos a tres como máximo	Evitar utilizar subgrupos pequeños (muestras menores a 5).	No: Edad + sexo + ocupación + # hijos Sí: Edad (agregada) + sexo + ocupación
Eliminar/Suprimir	Incluir solo los datos necesarios para que el estudio sea reproducible	Eliminar números identificadores, email, dirección, etc.

Tabla elaborada a partir de Best Practices: Sharing Human Subjects Data y la Guía básica de anonimización de AEPD.

k-anonimidad

Para garantizar que no se haya superado el umbral de riesgo de identificación, tras las técnicas previas, se pueden utilizar medidas como la k-anonimidad (o similares como l-diversity y t-closeness).

Es una técnica de protección de datos que busca garantizar que un individuo no pueda ser identificado de manera única dentro de un conjunto de datos anonimizados. Un conjunto de datos es k-anónimo si para cada combinación de identificadores indirectos existen al menos k-1 registros adicionales idénticos en esos atributos. Esto significa que la probabilidad de identificar a una persona a partir de esos datos es como máximo 1/k, reduciendo el riesgo de reidentificación.

Por ejemplo, si un conjunto de datos, cada combinación de valores de edad, código postal y género aparece en al menos cinco registros diferentes, se denominaría 5-anónimo, y evitaría la reidentificación de un individuo específico mediante esos datos.

Herramientas para anonimizar el set de datos

Amnesia: Utilizado para la seudoanimización de sets de datos tabulados. https://amnesia.openaire.eu/

Pydeface y mri_deface (FreeSurfer): Utilizado para eliminar/oscurecer estructuras faciales en datos fMRI https://pypi.org/project/pydeface/

ARX De-Identifier: https://arx.deidentifier.org/

Herramienta para crear identificadores únicos para participantes de los estudios: https://fitbir.nih.gov/content/global-unique-identifier

Herramientas para la anonimización: https://blog.gramener.com/10-best-data-anonymization-tools-and-techniques-to-protect-sensitive-information/

Aplicaciones para de-identificación en Investigación (Universidad Johns Hopkins): https://dataservices.library.jhu.edu/resources/applications-to-assist-in-de-identification-of-human-subjects-research-data/#image

ANEXO VII. Componentes del archivo README

Los archivos README incluyen información puede identificar y explicar su conjunto de datos a los usuarios, independientemente de si se accede a él o no.

Qué debe incluirse:

- Resumen del proyecto, experimento o intervención en los que se basa este conjunto de datos.
- Descripción de la estructura y el contenido del archivo.
- Definiciones de todas las variables, abreviaturas, códigos de datos faltantes y unidades.
- Enlaces a otras ubicaciones de los datos accesibles al público.
- Otras fuentes, si las hubiera, de las que se derivaron los datos.
- Cualquier otro dato que pueda influir en la reutilización o replicación de los datos recogidos.
- Se recomienda redactar en inglés, pero puede incluirse adicionalmente en otros idiomas, como el español, usando un nombre de archivo diferente (por ejemplo, README-es.txt).

Estructura del archivo

INFORMACIÓN GENERAL

1. Título del set de datos (Dataset Title)

   Resumir brevemente el contenido del conjunto de datos, contextualizado los procedimientos y resultados obtenidos.

2. Contacto (Contact person)

   Investigador/a de contacto

   Nombre:

   Filiación:

   Correo electrónico:

   ORCID:

3. Organización (Organization).

   Incluir el nombre y dirección del IIS o del centro

4. Descripción del proyecto (Project description)

   Proporcionar un resumen breve del proyecto asociado al set de datos, incluyendo su objetivo principal y finalidad.

5. Descripción de los datos y estructura de los archivos (Data description)

   Describir cómo están estructurados los datos y cómo podría utilizarlos otro investigador. Los usuarios del set de datos pueden desconocer terminologías específicas del campo, por lo que es recomendable explicarlas brevemente. Describir las relaciones entre los archivos de datos, los códigos de datos faltantes y otras abreviaturas utilizadas.

6. *Notas*
7. *Fecha de depósito de los archivos (Publication Date): YYYY-mm-dd*

8. *Fecha de creación de los archivos (Created Date): YYYY-mm-dd*
9. *Idioma (Language)*

AUTORÍA

1. Autores (*indicar autor/es principal/es y resto de investigadores*)

Principal Investigator:

Name:
Surname:
Phone (optional):
email:
ORCID:

Associate 1:

Name:
Surname:
Phone (optional):
email:
ORCID:

METODOLOGÍA

1. Metodología (Methods)
   Describir de forma clara cómo se recopilaron, procesaron y organizaron los datos. Indicar las técnicas empleadas (instrumentos, encuestas, scripts, criterios de limpieza) y cualquier decisión relevante que afecte a la interpretación.
2. Software
   Describir cualquier script, código o programa (por ejemplo, R, Python, Mathematica, MatLab), así como las versiones del software utilizado para ejecutar esos archivos. Si la relación del script con los archivos adjuntos no es obvia, es recomendable proporcionar información sobre el flujo de trabajo que se realizó para ejecutarlos.

PALABRAS CLAVE

1. Palabras clave (Keywords)

   Lista de palabras clave que describen el set de datos, indicando si provienen de un vocabulario controlado o libre.

INFORMACIÓN DE FINANCIACIÓN E IDENTIFICADORES DEL PROYECTO

1. Información sobre el proyecto (Grants)
   Incluir: Nombre, organismo financiador, código identificador, etc.

PUBLICACIONES RELACIONADAS

1. Publicación relacionada (Publication)
   Indicar la referencia bibliográfica completa, si aplica, incluyendo el DOI, URL o enlace a la versión publicada
2. Set de datos relacionado (Links to other research outputs and datasets)

Enlaces a las ubicaciones de conjuntos de datos relacionados accesibles al público

FICHEROS

1. Ficheros (Files)
   Incluir la relación de todos los ficheros que se incluyen en el set de datos. Utilizar la estructura siguiente para cada archivo incluido.
   File 1
   - *Name: [Nombre completo del archivo, Ej: sleep.csv]*
   - *Description: [Breve resumen del contenido, Ej: dataset with sleep variables and sociodemographic data]*
   - *SourceType: [el tipo de recogida de los datos, Elegir entre: Observed, Experiment, Computational]*
   - *DataSource: [Explicar cómo se han obtenido, Ej: Data collected from self-administered questionnaires]*
   - *FileFormat: [Formato del archivo, Ej: CSV]*
   - *Variables: [variable,descripcion,tipo;variable,descripcion,tipo, Ej: sleephr, horas dormido de media; despertares, nº de despertares nocturnos de media]*
   - *Idioma: español*

DERECHOS DE USO Y PRIVACIDAD

1. Licencias de uso (License)
   Indicar las licencias bajo las cuales se distribuye el set de datos, especificando permisos, restricciones y condiciones de uso.
2. Tratamiento de los datos (Data Treatment)
   Describir cómo se protegen los datos personales, indicando si se han anonimizado, se han eliminado identificadores o se aplican otras medidas de privacidad.

OTROS

1. Diccionario datos (Codebook)
   Listado detallado de todas las variables presentes en el set de datos. Para cada una de ellas indicar nombre, tipo de dato, descripción, posibles valores y unidades si aplica. En el caso de que exista un archivo adjunto con esta información, indicar su nombre y cómo acceder.
2. Cita del set de datos (How To Cite)

Bibliografía

Broman KW, Woo KH. Data organization in spreadsheets. The American Statistician. 2018 Jan 2;72(1):2-10.

Consorcio Madroño. *Readme.txt template*. e-cienciaDatos. 2025. [citado 2025 Sep 12] Disponible en: https://edatos.consorciomadrono.es/readme.xhtml

Corcho O, Sánchez González L, Simperl E; European Data Portal; Publications Office of the European Union. Report on Data Homogenisation for High-value Datasets [Internet]. Europa: Publications Office of the European Union; 2023 Dec 5. Report No.: OA-09-23-557-EN-N. ISBN: 978-92-78-43830-2. DOI: 10.2830/446704. Disponible en: https://data.europa.eu/en/publications/reports/report-data-homogenisation-high-value-datasets

Cunha-Oliveira T, Ioannidis JP, Oliveira PJ. Best practices for data management and sharing in experimental biomedical research. Physiological Reviews. 2024 Jul 1;104(3):1387-408.

Dryad Digital Repository. Best practices for creating reusable data publications [Internet]. Dryad; [citado 2025 Aug 29]. Disponible en: https://datadryad.org/best_practices

DCMI Usage Board. Dublin Core™ Metadata Element Set, Version 1.1: Reference Description [Internet]. Dublin Core Metadata Initiative; 2012 Jun 14 [citado 2025 Oct 21]. Available from: https://www.dublincore.org/specifications/dublin-core/dces/

Extract from the European Council Conclusions (2014/C 240/01.0001.01.ENG). Official Journal of the European Union. 24 July 2014; C-240:13. [Internet]. Disponible en: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.C_.2014.240.01.0001.01.ENG

FAIRsharing. Disponible en: https://fairsharing.org/ [citado 2025 Oct 15].

Guía básica de anonimización. Traducido por la Agencia Española de Protección de Datos (AEPD) de la guía elaborada por la Personal Data Protection Commission (PDPC) de Singapur. Publicado octubre 2022. Madrid: AEPD; 2022. PDF [Internet]. Disponible en: https://www.aepd.es/documento/guia-basica-anonimizacion.pdf.

Instituto de Salud Carlos III. Aspectos de seguridad en el manejo de datos sensibles. Madrid: Instituto de Salud Carlos III; 2023 [citado 2025 Oct 15]. Disponible en: https://impact.isciii.es/wp-content/uploads/2023/07/Aspectos-de-seguridad-en-el-manejo-de-datos-sensibles.pdf

Instituto de Salud Carlos III. Informe GT2: Plan de gestión de datos [Internet]. Ministerio de Ciencia e Innovación / ISCIII; 2023 [citado 2025 Oct 15]. Disponible en: https://www.isciii.es/documents/d/guest/informegt2_plangestiondatos_feb2023

Publications Office of the European Union. Metadata quality. In: *Data-Provider Manual*. [Internet]. [citado 2025 Aug 29]. Disponible en: https://dataeuropa.gitlab.io/data-provider-manual/metadata-quality/ GitLab

Publications Services, Stanford University Libraries. Data best practices and case studies [Internet]. Stanford (CA): Stanford University Libraries; [citado 2025 Aug 29]. Ultima actualización Sep 13, 2023. Disponible en: https://guides.library.stanford.edu/data-best-practices