

GUÍA PARA LA ELABORACIÓN DE UN PLAN DE GESTIÓN DE DATOS

Versión 5, marzo 2024

El Plan de Gestión de Datos (Data Management Plan) es un documento formal en el que se detalla y describe cómo se va a gestionar y tratar los datos de investigación recopilados o generados en un proyecto de investigación. Es un documento que se modifica y actualiza con el desarrollo del proyecto, aumentando su precisión y detalle. El objetivo de este documento es servir de guía a los investigadores que deben elaborar sus Planes de Gestión de Datos para las diferentes convocatorias de proyectos. Estos Planes de Gestión de Datos deben indicar las medidas tomadas por los investigadores para hacer sus datos FAIR, esto es, localizables, accesibles, interoperables y reutilizables.

Este documento incluye:

- Una descripción de los contenidos exigidos en Convocatorias Nacionales (ISCIII, AEI) y la plantilla para su elaboración.
- Una descripción de los contenidos exigidos en los Planes de Gestión de Datos en el marco Horizonte 2020 y la plantilla para su elaboración

ISCIII. Convocatoria I+D+I

Descripción de contenidos a incluir

Actualmente no se requiere en la convocatoria que se elabore un plan de gestión de datos completo, sino resumen ejecutivo en el que hay que abordar los siguientes elementos:

1. Tipología y formato de los datos

- Descripción del tipo de datos que se obtendrán:
 - Numéricos (Bases de datos, hojas de cálculo)
 - Texto (documentos)
 - Imágenes, audio, video o mezcla de ellos
- Indicar los formatos (extensión) para los archivos o conjuntos de datos (dataset).
- Identificar qué datos personales se recopilarán y utilizarán para su investigación y tratamiento de los mismos (pseudonimizados o anonimizados). Para la anonimización de los datos se pueden usar herramientas como <http://amnesia.openaire.eu> (recomendada por el ISCIII) o la de la Agencia de Protección de Datos (<https://www.aepd.es/descargas/herramienta-anonizacion-pdpc>)
- Especificar metodología para la recogida de datos, el software utilizado (p.ej. RedCap) y los controles de calidad que se utilizarán.
- Indicar, si las vamos a utilizar, que tenemos autorización para el uso de bases de datos públicas o privadas.

2. Procedimiento para su acceso

- Personas del equipo que tendrán acceso
- Uso de plataformas de bases de datos (p.ej. RedCap) y si se almacenan en servidores protegidos, en la nube o en ordenadores propios. Incluir si se harán copias de seguridad durante el proyecto y su periodicidad.
- Especificar ciertas medidas de seguridad si se van a enviar datos:
 - No usar cuentas de Gmail, Hotmail o similares. Utilizar siempre el correo corporativo
 - Los datos deben estar cifrados y con contraseñas de calidad
 - Los datos deben enviarse pseudoanonimizados, eliminando todo identificador indirecto o directo no necesario (p.ej. indicar año de nacimiento en lugar de la fecha completa)

3. Titularidad de los datos

- Dependerá del origen del dato:
 - Titularidad del paciente: Datos obtenidos de sus pruebas o de la historia (Analíticas, Pruebas histopatológicas, etc.), así como todo dato extraído mediante consentimiento informado.

- Titularidad de la institución: Datos elaborados a partir de datos puros o estudios de experimentación animal. Si se utilizan muestras de pacientes, la titularidad será compartida (paciente e institución)
- En el caso de datos extraídos de fuentes externas (papers o repositorios) o datos de bases privadas o públicas (secuencias genómicas, proteómicas, etc.), la titularidad corresponderá a la entidad de origen.

4. Repositorio donde se preservarán

- Se han de depositar en los repositorios todos los datos que subyacen a la investigación, es decir, los datos brutos generados o producidos en el transcurso de la investigación.
- Se deberán publicar junto a los artículos científicos los datos finales que sean necesarios para garantizar la verificación y reproducibilidad de los resultados presentados.
- En el depósito y publicación se deberá tener en cuenta la protección de los datos personales y protección ante el uso de terceros no autorizados, así como la privacidad y confidencialidad de los datos.
- Desde el IBSAL recomendamos registrarlo en Zenodo (<https://zenodo.org/>), repositorio apoyado por la Unión Europea, ya que se cuenta con una **comunidad del IBSAL** para tal fin. Se pueden registrar en repositorios institucionales (p.ej. www.gredos.usal.es) o en el repositorio del Instituto de Salud Carlos III (www.repisalud.isciii.es), cuando esté disponible. Especificar en este apartado el nombre del repositorio seleccionado.
- Determinar quién será el responsable de desarrollar, archivar en el repositorio y actualizar el Plan de Gestión de datos (Data Manager).
- Incluir el tipo de licencia que se utilizará para los datos y archivos (Creative Commons license). Una herramienta de ayuda para seleccionarla es <https://chooser-beta.creativecommons.org/>

5. Procedimiento de los requisitos éticos

- En el caso de que se registren datos personales, explicar cómo se garantizará el cumplimiento de la legislación sobre datos personales.
- Especificar las condiciones éticas o legales específicas que condicionen la disponibilidad de los datos y/o su reutilización si son datos protegidos o protegibles por propiedad intelectual o industrial.

Plantilla para su elaboración

Tipología y formato de los datos

Los datos clínicos y demográficos se obtendrán de ([historia clínica electrónica del hospital, entrevista estructurada, etc.](#)). Las variables necesarias para el estudio están descritas en el apartado de metodología. Los datos del estudio se recopilarán y administrarán utilizando las herramientas de captura de datos electrónicos del sistema ([REDCap \(Research Electronic Data Capture\) o la que corresponda](#)). Esta plataforma permite la trazabilidad, exportación en formatos como CSV, R y XML, facilitando su integración con paquetes estadísticos comunes (SPSS, SAS, STATA, R). La plataforma garantizará la pseudoanonimización de los datos, cumpliendo con los principios FAIR. La descripción detallada de los datos, junto con la metodología, se incluirá en un fichero readme.txt junto con el dataset, y se acompañará de metadatos estandarizados.

Acceso

Solo el personal investigador autorizado tendrá acceso a ([donde se recojan los datos](#)) con una cuenta personal. Se implementarán medidas de seguridad específicas para evitar la identificación y el acceso no autorizado de terceros. La descarga de datos estará restringida a IP y a quien sea asignada esta función. Se realizarán copias de seguridad de la base de datos de forma ([mensual, semestral, anual](#)), las cuales se almacenarán en dispositivos protegidos. Se llevarán a cabo los pertinentes controles de calidad como la calibración, repetición de mediciones y validación de entrada de datos. Los análisis estadísticos se harán sobre datos totalmente anonimizados.

Titularidad de los datos

Según la propuesta de reglamento del parlamento europeo y del consejo sobre el espacio europeo de datos sanitarios, el titular de la información de datos sanitarios es el participante del estudio. El responsable de los datos es el organismo donde se realiza el estudio ([SACYL, USAL, CSIC, etc.](#)) y el promotor/investigador principal es el encargado de su tratamiento en el ámbito de la investigación. Los datos objeto de tratamiento se almacenarán en dispositivos bajo el control del organismo responsable, cumpliendo los requisitos relevantes en las dimensiones de integridad, autenticidad, confidencialidad, trazabilidad y conservación de los datos. La salida de datos fuera del control técnico del organismo responsable será excepcional y requerirá justificación previa del investigador principal, autorizada por el responsable.

Repositorios donde se preservarán los datos

Los datos incluidos en los repositorios serán anonimizados, utilizando si es necesario la herramienta amnesia y se depositarán en el repositorio digital abierto Zenodo, dentro de la Comunidad del IBSAL creada para este propósito. Un investigador designado dentro del equipo ([si se sabe quién será el responsable, especificar su nombre](#)) que será el responsable de desarrollar y mantener el Plan de Gestión de Datos, entregándolo en los primeros 6 meses del proyecto. Los datos se almacenarán en un formato exportable (.csv) y se publicarán bajo la licencia Creative Commons ([CC-BY](#)), asignándoles también un DOI y url (handle) para aumentar su reutilización.

Procedimientos éticos y legales

El proyecto contará con el informe previo favorable del Comité Ético de Investigación con medicamentos (CEIm) del Área de Salud de Salamanca. El desarrollo del proyecto se llevará a cabo conforme a las normas éticas vigentes, la Declaración de Helsinki y la legislación de protección de datos (Ley 3/2018 y Reglamento Europeo 2016/679). Se garantizará la confidencialidad y seguridad de la información, eliminando los datos sin anonimizar y restringiendo el acceso en caso de patentes.

CONVOCATORIAS EUROPEAS

Los investigadores que soliciten proyectos en las convocatorias de Horizonte Europa deben cumplir, respecto a los datos de investigación, una serie de requisitos que deben estar recogidos en el Plan de Gestión de Datos. Este documento se debe entregar durante los 6 primeros meses del proyecto como un deliverable del mismo.

Los proyectos participantes deben cumplir con 2 pasos (<https://bibliotecas.csic.es/node/195>):

- Depósito de los datos, incluidos sus metadatos, en un repositorio
- Los proyectos deben adoptar medidas que permitan el acceso, la explotación, la reproducción y la difusión de los datos usando, por ejemplo, las licencias Creative Commons (CC-BY o CC-0). También se debe informar, a través del repositorio, sobre las herramientas necesarias para validar los resultados.

La estructura e información a incluir en un Plan de Gestión de Datos, de acuerdo con el documento Horizon Europe Data Management Plan Template (https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/temp-form/report/data-management-plan_he_en.docx) y el DMP Checklist (<http://www.dcc.ac.uk/resources/data-management-plans>) es la siguiente:

2.1. Información del Proyecto

Nombre e identificación del proyecto, descripción del mismo, Institución, Agencia de financiación, investigador principal e identificador (ID), datos de contacto, y versiones del Plan de Gestión de Datos.

2.2. Resumen de datos

Describir los datos que se utilizarán en el proyecto, incluyendo el tipo y el formato, la finalidad, el tamaño, origen y utilidad de los datos.

- Reutilización de los datos: Indicar si los datos son reutilizados y la finalidad de su uso, así como la fuente de donde se han extraído. Tratamiento de derechos de autor respecto a datos reutilizados.
- Tipo y formato de datos: Explicar qué tipo de datos generará el proyecto, (observacionales, experimentales, simulación, derivadas o compiladas, de referencia, otros) y el formato de los mismos (jpeg, pdf, docs., xlsx...).
- Justificar el uso de determinados formatos. Por ejemplo, las decisiones pueden basarse en la experiencia del personal de la organización, la preferencia por formatos abiertos o las normas aceptadas por los repositorios de datos.
- Propósito de los datos y relación con el objetivo del proyecto: Indicar el propósito de la recogida y su utilización, así como la relación de éstos con los objetivos del proyecto.
- Tamaño de los datos: Especificar el tamaño esperado de los datos que se generarán. Indicar el volumen del dataset en MB/ GB/ TB y el crecimiento gradual.
- Origen de datos: Indicar la procedencia de los datos. Si son generados, indicar la metodología, y si son reutilizados, indicar la fuente y los autores.
- Utilidad de los datos: Describir la utilidad para futuras investigaciones
- Indicar cómo se organizarán los datos durante el Proyecto, como las versiones de control y estructura de las carpetas/archivos.

2.3. Datos FAIR

El PGD debe garantizar que los datos cumplen con los principios FAIR (Findable, Accessible, Interoperable, Re-usable), es decir disponen de los requisitos necesarios para que sean Encontrables, Accesibles, Interoperables y Reutilizables.

2.3.1. Hacer que los datos se puedan encontrar

Incluir la información sobre Metadatos para que los datos sean localizables. Indicar:

- Identificadores persistentes (como el DOI). Indicar claramente si se proporcionan para todos los conjuntos de datos, conjuntos de datos individuales, colecciones de datos o subconjuntos, y cuáles. Si no se van a utilizar, explicar por qué.
- Palabras claves que optimicen las posibilidades de reutilización.
- Describirlos con metadatos enriquecidos
- Nomenclaturas utilizadas.
- Número de versiones.
- Utilización de metadatos estándares (p.ej. DDI, TEI, EML, MARC, CMDI). En caso de no existir estándares de metadatos en el área del proyecto, describir qué tipo de metadatos se crearán y cómo.

2.3.2. Hacer que los datos sean accesibles

Se especificarán las características del Repositorio donde se depositarán los datos, así como la accesibilidad de los datos y su descripción con metadatos.

- Indicar el repositorio donde se depositarán los datos. En el caso del IBSAL se realizaría en la comunidad del IBSAL en Zenodo ([LINK](#))
- Los conjuntos de datos están identificados con DOI y url.
- Indicar qué datos producidos y/o utilizados en el proyecto estarán en abierto. Si algún conjunto de datos no se puede compartir o se debe compartir con restricciones, hay que explicar las razones legales y contractuales de éstas e indicar cómo se proporciona el acceso.
- Indicar si los metadatos estarán en acceso abierto
- Indicar métodos o herramientas de software para acceder a los datos (p.ej., scripts, códigos o algoritmos específicos desarrollados durante el proyecto, versión del software)

2.3.3. Hacer que los datos sean interoperables

Los datos producidos en el proyecto deben ser interoperables, permitiendo el intercambio y la reutilización de los mismos entre investigadores, instituciones, organizaciones, países etc.

- Indicar vocabularios y estándares que se usarán.
- Indicar las normas, formatos o metodologías que se utilizarán para permitir el intercambio y la interoperabilidad de los datos.

2.3.4. Aumentar la reutilización de los datos

Documentar la procedencia de los datos y proporcionar la información necesaria para validar los datos y facilitar su reutilización, por ejemplo, con los archivos readme. Proporcionar las licencias de uso de los datos.

- Indicar qué tipo de licencia se ha utilizado para permitir la reutilización más amplia posible.
- Indicar cuándo estarán los datos disponibles. Si es necesario un tiempo de embargo para que los datos estén disponibles para su reutilización, hay que indicar el motivo que causa este embargo.
- Indicar si los datos producidos y/o utilizados en el proyecto pueden ser utilizados por terceros. Si la reutilización de algunos datos está restringida, indicar por qué.
- Indicar el tiempo que estarán los datos disponibles para ser reutilizados.
- Describir los procesos de comprobación de la calidad de los datos. Puede incluir procesos como la calibración, la repetición de muestras o mediciones, la captura normalizada de datos, la validación de la introducción de datos, la revisión por pares de los datos o la representación con vocabularios controlados.

2.4. Otros resultados de la investigación

Se consideran aquí qué cuestiones relativas a los datos FAIR pueden aplicarse a la gestión de otros productos de la investigación generados o reutilizados en el proyecto como, por ejemplo: softwares, flujos de trabajo, protocolos, nuevos materiales, muestras, etc. y cómo se gestionarán y compartirán, o se pondrán a disposición para su reutilización, de conformidad con los principios FAIR.

2.5. Asignación de recursos

Se especificarán las asignaciones tanto de recursos humanos como económicos.

- Describir las funciones y responsabilidades relativas a la gestión/conservación de datos, por ejemplo, la captura de datos, la producción de metadatos, la calidad de los datos, el almacenamiento y las copias de seguridad, el archivo de datos y la puesta en común de datos. En la medida de lo posible, indique la/s persona/s responsables.
- Indicar quién es el responsable de aplicar el plan de gestión de datos, de su revisión y modificación.
- Indicar cuáles serán los costes para hacer que los datos sean FAIR (por ejemplo, costes directos e indirectos relacionados con el almacenamiento, el archivo, la reutilización, la seguridad, etc.) y quién será responsable de la gestión de los datos.

2.6. Seguridad de los datos

Garantizar que los datos se almacenan de forma segura en repositorios de confianza para su preservación y conservación a largo plazo.

- Describir dónde se almacenarán los datos y se realizará una copia de seguridad de los mismos durante el proyecto y con qué frecuencia se realizará la copia de seguridad. Se recomienda almacenar los datos al menos en dos ubicaciones separadas y el uso de un almacenamiento sólido y gestionado con copia de seguridad automática. No se recomienda almacenar los datos en ordenadores portátiles, discos duros independientes o dispositivos de almacenamiento externo como memorias USB.
- Explicar cómo se recuperarán los datos en el caso que hubiera un incidente.
- Explicar quién tendrá acceso a los datos durante la investigación y cómo se controla el acceso a los datos, especialmente en colaboraciones con otros grupos.
- Tenga en cuenta la protección de datos, sobre todo si son sensibles (p.ej., si contienen datos personales o secretos comerciales). Describir los principales riesgos y cómo se gestionarán.
- Explicar qué políticas institucionales de protección de datos existen.

2.7. Ética

Se presentan todas las cuestiones éticas o legales que puedan tener un impacto en el intercambio de datos. También pueden tratarse en el contexto de la revisión ética.

- Describir los posibles problemas éticos durante la recogida, el almacenaje, el procesamiento y el archivo de datos, junto a los procedimientos de aprobación ética relacionados con el proyecto.
- Incluir el consentimiento informado para el intercambio de datos y la conservación a largo plazo de los cuestionarios que tratan con datos personales.
- Estudiar la posibilidad de anonimizar los datos personales para conservarlos y/o compartirlos o de pseudanonimizarlos (la diferencia es que este proceso es reversible).

2.8. Otros aspectos

Se describe si se van a utilizar, otros procedimientos nacionales, financiadores, sectoriales o departamentales para la gestión de datos.

Documentos de ayuda

Hay varias herramientas que nos pueden facilitar la creación del PGD:

- Data Stewardship Wizard (<https://ds-wizard.org/>)
- Argos (<https://argos.openaire.eu/splash/index.html>)
- DMP Online (<https://dmponline.dcc.ac.uk/>)
- DMP Tool (<https://dmptool.org/>)

- PGD Online (<https://pgd.consorciomadrono.es/home>)

Data Management Plan - Template

Action Number: [insert project reference number]

Action Acronym: [insert acronym]

Action title: [insert project title]

Date: [insert date]

DMP version: [insert DMP version]

1. Data summary

1.A What is the purpose of the data generation or re-use and its relation to the objectives of the project?

1.A a) Description

Briefly explain the purpose of the data generated and/or collected and its relation to the project objectives.

Drafting examples:

- To facilitate data exchange within the project discipline.
- Experimental data will be collected in order to fulfil the commitments established with the funding agency.
- To meet the project objectives and in peer-reviewed conferences and publications.
- To collect research data under the project and make it available to support the credibility and enhance the quality of scientific publications based on such data.
- To provide data for future research projects to continue with the work. This will ensure the lasting impact of this EC-funded project beyond the project period.

1.A b) Real example

Ex. 1 *The data will originate from measurements, calibrations, comparisons and validations. It will be used in meeting the project's objectives and in conference and peer-reviewed publications.*

Experimental data will be collected by the consortium in order to meet objectives 1 - 4. Measurement and calibration data will result from objectives 1 and 3 and comparison and validation data from objectives 2 and 4. Data from questionnaires and market surveys will be used to support end-user uptake (objective 5).

Ex. 2 *Collecting and making available the data of the analysis of superconducting materials to support the credibility and raise the quality of the scientific publications based on those data. Ease the exchange of data within the Consortium and promote the distributed characterization of samples with different methods. Permit follow-up projects and further generations of students continuing the work to build upon existing data sets, to validate the results and to document the improvement of materials and production techniques in a verifiable manner. This approach will ensure a durable impact of this EC funded project beyond the project period.*

The objective of the project is to advance the performance of superconducting wires and at a later stage thin films by gaining a better understanding of the material behavior, the influence on the production techniques on the performance and to elucidate performance limitations (e.g. quality factor for superconducting thin films on substrate, current limits in wires under high-magnetic field conditions). Managed collection and publication of the data shall help establishing a durable library of results that can help documenting the performance evolution across several years and to permit other researchers validating the results independently.

1.B What types and formats of data will the project generate or re-use?

1.B a) Description

Description of the content and scope of the data. Research data are generated for various reasons and through various processes, and may be of the following types:

- Observational: data captured in real time (neuroimages, sample data, sensor data, survey data, etc.).
- Experimental: data captured by laboratory equipment (gene sequences, chromatograms, magnetic field data, etc.).
- Simulation: data generated from test models (climate, mathematical, economic, etc.).
- Derived or compiled: data that are reproducible but difficult to reproduce (text and data mining, 3D models, compiled databases, etc.).
- Reference: conglomerated datasets (databases of gene sequences, chemical structures, spatial data portals, etc.).
- Others

Format of the data (text, numeric, image, etc.) must also be indicated.

1.B b) Real example

Ex. 1 *The majority of the data will be in ASCII (American Standard Code for Information Interchange) data files, eg comma separated variable (CSV) format, which can be imported into rich-text files for word-processing or into spreadsheets. If specialised software is used, then information about free readers will be provided. Data will be generated in the following formats:*

- Graphics: jpeg, odg, pdf, png, pttx
- Tables: odsu, opj, xlsx
- Text: docx, pdf, txt
- Other: nb, cpp

Ex. 2 The openly accessible data will be the comprehensive result data sets of characterized samples that are used to create the figures and plots in scientific publications, such that other researchers can compare their results easier and such that further results including historic data can be produced quicker. The data are value tables in Open Document Spreadsheet format (.ODS) for limited amounts of data with typed columns. For larger quantities of numeric data, UTF-8 encoded, comma separated value in textual format files (.CSV) with column value and data format description (FORMAT.TXT) will be used. In addition, images and raw measurement data files as provided by the measurement instruments will be stored on a project-internal data storage platform. Data files and images will be included in the open data sets. Proprietary raw data delivered by the measurement instruments will not be published. For all published files, a document record and change track will be included (author contact information, status, version, change reason and date, description of contents, title, origin of the data including a brief description of the measurement and/or experiment setup) in a separate metadata file for each characterization action called METADATA.ODS.

1.C Will you re-use any existing data and what will you re-use it for? State the reasons if re-use of any existing data has been considered but discarded

1.C a) Description

If you reuse a dataset, specify the source from which it was extracted for example from a relevant repository. If purchasing or reusing existing data sources, explain how issues such as copyright and IPR have been addressed.

When creating new data sources, explain why existing data sources cannot be reused.

1.C b) Real example

Ex. 1 Some of the project's tasks will use existing data in hdf, txt and xlsx formats. These data will be used in the validation of the project's results.

Ex. 2 Existing data from ongoing R&D projects in the scope of the <AcronymProject> study on superconducting wires and thin films will serve as a basis for the data files.

Ex. 3 Selected, existing images and data from the databases of the partner museums (<Partner1>, <Partner2>, <Partner3>...) will be used in specific tests, such as the storage tests in WP6. The final kind of data that will be created is that which is information in project deliverables, which must be preserved, made accessible and passed on to subsequent persons working in <AcronymProject>.

1.D What is the origin/provenance of the data, either generated or re-used?

1.D a) Description

If the data are generated within the project, state the source of the data.

If the data are collected, state the source from which they were extracted.

1.D b) Real example

Ex. 1 The existing data will originate from several sources, which will include: partner's pre-existing data, data from the scientific literature, real-world measurement data and data from simulation experiments. The data collected from domestic properties will remain confidential and will not be included in the repository.

Ex. 2 The data stem from experiments and measurement campaigns performed by the ESRs and their colleagues at the beneficiary institutes: 1. Phase A: Superconducting wires and tapes: <Partner1>, <Partner2>, <Partner3>... 2. Phase B: Superconducting thin films: <Partner1>, <Partner2>, <Partner3>....

Ex. 3 These data have been digitised in diverse earlier projects.

1.E What is the expected size of the data that you intend to generate or re-use?

1.E a) Description

State the approximate volume of the datasets. Consider the implications of data volumes in terms of storage, backup, cost and access. Estimate the volume of data in MB/GB/TB and how this will grow to make sure any additional storage and technical support required can be provided.

1.E b) Real example

Ex. 1 The expected size of the data is not currently known, but it is likely to be <10 GB with individual files being ≤1 MB.

Ex. 2 The size of the data is today not known. Initial experience with storing results from different kind of measurements will permit revising this initial data management plan. The main relevant data sizes will stem from images such as microscopic sample characteristic that are stored in high-resolution bitmap format. However, the total data set size for a single sample characterization is expected to be in the order of tens of MB only.

Ex. 3 The size of the data handled by <AcronymProject> is quite small, such as less than 10 GB, except in the tests of the data infrastructure in WP6, where the project needs experience of managing large volumes of data, as explained above.

1.F To whom might your data be useful ('data utility'), outside your project?

1.F a) Description

State the group/s who may be interested in the data.

1.F b) Real example

Ex. 1 The data will be suitable for use by other research groups working on the following topics: biogas, biomethane, energy gases. It will also be useful for standards committees including ISO/TC193/SC1/WG25 Biomethane Working Group, ISO/TC 158 Analysis of Gases and regulators.

Ex. 2 Within the Consortium:

The data sets will be shared within the consortium as the working baseline to produce the scientific publications, to verify and validate the results through repeated experiments at different locations and as a baseline for a comprehensive documentation of the superconducting material performance evaluation in the scope of the world-wide Future Circular Collider Technology R&D program.

Beyond the Consortium:

The data can be used by independent researchers to understand better the contents and conclusions of the scientific publications, which base their findings on the data. Furthermore, independent researchers can use the files to produce figures and publications, showing comparisons of their own results and the <AcronymProject> results. Scientists can also use the data files to repeat the experiments and measurements to verify and validate the <AcronymProject> research. Finally, the data sets may also be used by scientific writers and the press to produce high-quality infographics, demonstrating the impact potentials of the technology.

Ex. 3 The data from these limited pilots will be useful for users and institutions who may be considering similar technologies in their digitisation and data management work. This applies in particular to the experiments carried out by WP6, but also the others. In particular, the digitised data from the experiments in WP3 will make apparent the quality of the digitisation results achieved with the new technologies. The data in the experiments of WP5 will be useful for the museums.

2. FAIR Data

2.1 Making data findable, including provisions of metadata

2.1.A Will data be identified by a persistent identifier?

2.1.A a) Description

Explain how the data and metadata are assigned to a globally unique and eternally persistent identifier (DOI, Handle...).

2.1.A b) Real example

Ex. 1 The CORA – Research Data Repository (RdR) (<https://dataverse.csuc.cat>) assigns DOIs for persistent identification and identifiability of the dataset.

Ex. 2 The repository assigns Handle/DOIs for persistent identification and citability of the dataset.

2.1.B Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

2.1.B a) Description

Rich metadata provided: The metadata should document how the data were generated, under what license and how they can be re-used. Also, metadata helps to discover the data and provide the context for proper interpretation by other researchers.

Metadata created and standards: State the metadata standards that will be used. We recommend using metadata standards that are specific to the discipline. Consult metadata standards.

If metadata standards are not used, state what metadata will be generated (manually or automatically) and how.

2.1.B b) Real example

Ex. 1 The metadata standard used to describe the dataset will be the Dublin Core Schema, as it is a flexible and common used standard and is also the one adopted by the European OpenAIRE repository.

Ex. 2 Metadata are created manually by depositors in the deposit form at the repository.

Ex. 3 (1) The data are expected to be provided in ANSI SQL, XML or text (ASCII) format. For this dataset, data citation and metadata practices derived from the community will be considered.

(2) There are no standards for these logs. A possible solution is project servers such as AAA servers. In this case, the logs would include the attributes defined by "project".

Ex. 4 Each file associated with data will be accompanied with unique specified metadata to allow ease of access and re-usability. Below, the form to be followed is presented.

Ex. 5 Standards such as the Dublin Core and ISO/IEC 11179 Metadata Registry (MDR), which addresses issues in the metadata and data modelling space, will be considered.

Ex. 6 There are many different metadata standards for many different types of data and it may not be possible to find one that fits all purposes. Therefore, a pragmatic and feasible approach is to agree on a common and minimal catalogue metadata schema for those datasets that are published in public catalogues and data repositories and to use data-type specific schema extensions, if necessary. In

general, the Zenodo deposition metadata domain model which is based on DataCite's metadata schema minimum and recommended terms will be used for open data generated by the project and deposited in an appropriate repository.

2.1.C Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

2.1.C a) Description

State how content search keywords will be created to optimize retrieval and reuse.

2.1.C b) Real example

Ex. 1 Data has to be findable easily, rapidly and identically. Therefore, exact and standard measures have to be used to identify the data sets. This can include the definition and use of naming conventions, search keywords, version numbers, metadata standards and standard data identifiers.

Ex. 2 All open project results deposited in a repository will provide search keywords together with their metadata. Keywords for open data will be selected from controlled vocabularies that are suitable for the specific type of data.

2.1.D Will metadata be offered in such a way that it can be harvested and indexed?

2.1.D a) Description

Metadata should be provided structured using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) in such a way as to allow exchange with other repositories. At the same time, the metadata provided should be as detailed as possible to allow it to be indexed so that the data is searchable and retrievable

2.2 Making data accessible

2.2.A Which data produced and/or used in the project will be made openly available as the default?

2.2.A a) Description

Description of whether and how data will be shared, including access procedures, embargo periods (if any), and definition of whether access will be wide open or restricted to specific groups. If some cannot be made openly available, you must justify why.

2.2.A b) Real example

Ex. 1 All data produced by the experiments of WP3, WP4, WP5, and WP6, which has been described above, will be made openly available. This is any imagery and results of automatic or computer-assisted human interpretation of the data, which can be seen in the imagery. This does not mean that also the details of the equipment used, and algorithms used in the interpretation will be made openly available, as these may contain proprietary information. In Zenodo, the option exists to provide open access, embargoed access, closed access.

Ex. 2 All of the data associated with scientific publications will be made openly available as the default unless there is a specific reason not to publish the data. Datasets which cannot be shared – voluntary restrictions Other data may be made available on a case-by-case basis if it is relevant for third parties.

The following data will not be made publicly available:

- Data obtained with the permission of third parties, but the third parties have not agreed to make the data publicly available.
- Data that discloses the identity of a manufacturer.
- Data that compromises the protection of a partner(s) intellectual property. The level of data made available will also be considered, for example, pre-processed data will not be provided unless there is a clear reason for doing so.

Datasets which cannot be shared - legal and contractual reasons All of the data from the project will be made available, with the exception of market or customer survey data, which are commercially sensitive and cannot be shared.

2.2.B How will the data be made available (e.g. by deposition in a repository)

2.2.B a) Description

Describe how the data will be shared, i.e., who will have access to the dataset. You can create a procedure to temporarily make the data accessible to other group members, project partners, and the general public. You should state whether the data will be open access and in what reasonable period. One possibility is to offer them together with the publications. If embargo periods are required, this is where you need to specify them.

2.2.B b) Real example

Ex. 1 The data will be deposited in the storage systems which will be tested by WP6, as appropriate (<Repository1>, <Repository2>, <Repository3>). Links from <AcronymProject> website will be provided to these storage systems. By their service definition, the data stored at <Repository1> remains permanently available. Permanent access to the data on national <Repository2> and <Repository3> tests is not foreseen. Data from the digitisation pilots may remain permanently available if published on <Repository4>. These arrangements will be revisited after the data from the pilots has been created.

Ex. 2 Once processing, quality control, organisation, analysis and publication are complete, the data will be made accessible by deposition in open access repositories (eg Zenodo).

2.2.C What methods or software tools are needed to access the data

2.2.C a) Description

Include any technical requirements for access to and reuse of data.

2.2.C b) Real example

Ex. 1 All the data generated will be accessible via commonly used software: MS Office, Open Office, Adobe Reader, Image Viewer, etc. If any specific software is used, the necessary documentation relating to that software will be provided to

facilitate access to the technical data. For software developed under the project, the source code will be deposited in the repository.

Ex. 2 *Where possible, standard publicly available software will be used, but if specialised software tools are developed a short text file (e.g. ASCII) will be provided with the data file to explain the required software.*

2.2.D Is documentation about the software needed to access the data included?

2.2.D a) Description

You must also include the documentation on the software that is needed to access the data.

2.2.D b) Real example

Ex. 1 *If accessed through the API, documentation will be needed.*

Ex. 2 *Standard publicly available software will be used where possible, but if specialist software tools are developed, i.e. created within Matlab, a short text file (e.g. ASCII) will be provided with the data file to explain the software required.*

2.2.E Is it possible to include the relevant software (e.g. in open source code)?

2.2.D a) Description

In case it's specific software, it includes software if possible. For example, in the open source code.

2.2.D b) Real example

Ex. 1 *Any such software has already been released by the providers of these storage systems.*

Ex. 2 *The majority of the software programmes are available as commercial products or as freeware. For the software developed in the project, the source code will be deposited in the repository (eg Zenodo).*

2.2.F Where will the data and associated metadata, documentation and code be deposited?

2.2.F a) Description

State the repository in which the data and associated metadata, documents and code will be stored. It can be the same repository or different repositories for the different types of content, for instance, code could be deposit in a specific repository for code.

It is important to use a repository that provides permanent links (DOI, handle) to data in order to facilitate findability and citation.

2.2.F b) Real example

Ex. 1 *The data will be deposited in the storage systems which will be tested by WP6, as appropriate (national OSC, EUDAT, Zenodo). Links from <AcronymProject> website will be provided to these storage systems.*

Ex. 2 The data and associated metadata, documentation and code will either be deposited in the open access repository called Zenodo or in Open Access Repository (<ExampleURL>).

2.2.G Have you explored appropriate arrangements with the identified repository?

2.2.G a) Description

State if you have explored appropriately which are the requirements of the identified repository.

2.2.G b) Real example

Ex. 1 We have already explored the appropriate arrangements with the national cloud services in Finland (CSC), EUDAT through the work of <AcronymProject> pilot, and Zenodo through the work of the <Disciplinary> Literature Repository community.

Ex. 2 Yes, Open Access Repository is functional and it correctly labels datasets with a metadata scheme that is compatible with DataCite).

2.2.H Specify how access will be provided if there are restrictions on use

2.2.H a) Description

In case public access to data is restricted for any justified reason please specify if data would be accessible to an individual partner, to all partners or under request. Specify procedures of how to request access to restricted data and under which conditions it would be granted. Moreover, specify if restrictions will be lifted after a period of time.

2.2.H b) Real example

Ex. 1 There are no restrictions on use, except when CC BY-NC license has been chosen. <AcronymProject> should address question of sensitive data (e.g. location of protected plants), but <AcronymProject> will avoid working with any sensitive data. If personal data is received in questionnaires, which <AcronymProject> will receive, such data shall be anonymised before making available outside the project.

Ex. 2 There are no restrictions on the use of the published data, but users will be required to acknowledge the consortium and the source of the data in any resulting publications.

2.2.I Is there a need for a data access committee?

2.2.I a) Description

Specify why or why not it is necessary for a data access committee.

2.2.I b) Real example

Ex. 1 Because of the small scale of these experiments, there is no need for a data access committee.

Ex. 2 This consortium will have a data access committee. Their remit will be to select the data that will be openly accessible on a case by case basis. Ethical aspects and data security, including intellectual property requirements, will be considered. If

necessary, some or all of a potential publication's data will be withheld. This will be decided in consultation with the relevant partner(s).

2.2.J Are there well described conditions for access (i.e. a machine readable license)?

2.2.J a) Description

Describe which are the conditions for access defined by the repository you have chosen (you can also indicate the URL where the information comes from). For example: a machine-readable license.

2.2.J b) Real example

***Ex. 1** The Creative Commons licenses supported by the GBIF will be used. These include CC0, CC-BY, and CC BY-NC (see <ExampleURL>). Zenodo supports a large array of widely used as well as domain specific, machine-readable licences. The owner of the data will determine which of these licenses will be used when data is posted on <AcronymProject> repositories. However, it is the project's recommendation to choose CC0 for data and CC-BY for media and avoid CC-BY-NC which has issues in some national jurisdictions.*

***Ex. 2** Yes, Zenodo provides well-described conditions for access (see <http://about.zenodo.org/policies/>).*

2.2.K How will the identity of the person accessing the data be ascertained?

2.2.K a) Description

Describe the procedure established by the repository to determine the identity of the person accessing the data.

2.2.K b) Real example

***Ex. 1** Identity of the person accessing the data will not be directly ascertained. However, we expect users to follow the standard norms of scientific citation and use of the data in this context will be tracked through scientific citation.*

***Ex. 2** Users are required to register to use the repository.*

2.3 Making data interoperable

2.3.A Specify what data and metadata vocabularies, standards or methodologies you will follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

2.3.A a) Description

Explain what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability. The data interoperability of the project allows the exchange and reuse of data between researchers, institutions, organizations, countries, etc. Adhere to the standards of formats that are, as far as possible, compatible with open programs and applications.

2.3.A b) Real example

Ex. 1 The data produced in the project will be interoperable as the datasets will adhere to standardised formats: ASCII, txt, csv, xml, tiff. If MS Office, pdf viewer or image viewer cannot be used, a text (ASCII) file will be provided with the dataset that explains where a free reader can be obtained.

Ex. 2 Other types of data have been registered following internal codifications, clearly specified within the file.

2.4 Increase data reuse (through clarifying licenses)

2.4.A How the documentation needed to validate data analysis and facilitate data re-use will be provided?

2.4.A a) Description

You must also include the documentation that is needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.).

2.4.A b) Real example

Ex. 1 The data will be accompanied by a readme file for deposit with complementary information to aid the interpretation and re-use of the data.

2.4.B How the data will be licenced to permit the widest reuse possible, in line with the obligations set out in the Grant Agreement?

2.4.B a) Description

If the data are made available to other researchers and the general public, you need to specify what degree of reuse is allowed. This level of reuse will be marked by the establishment of licenses. The EC proposes the use of Creative Commons CC-BY or CC0 licences, but there are others.

2.4.B b) Real example

Ex. 1 The data will be licensed under Creative Commons CC-BY licences and will continue to be re-usable after the end of the project by anyone interested in the project, with no access or time restrictions

2.4.C Are the data produced and/or used in the project useable by third parties, after the end of the project? If the re-use of some data is restricted, explain why.

2.4.C a) Description

In principle, the data should be made available to other researchers and the general public with the fewest possible restrictions. However, there may be several reasons for not sharing them: ethical reasons, protection of personal data, the involvement of intellectual and/or industrial property rights, commercial interests, etc. You must specify the reasons why a dataset will not be shared.

2.4.C b) Real example

Ex. 1 The full dataset will be confidential and only the members of the consortium will have access to it. Furthermore, if it is decided to make specific portions of it (e.g.

metadata, statistics, etc.) widely open access, a data management portal will be created that should provide a description of the dataset and link to a download section. Of course, these data will be anonymized so as not to have any potential correlation and identification of the ethical issues with their publication and dissemination.

Ex. 2 Each archived data set will have its own permanent repository ID and will be easily accessible. We expect most of the data generated to be made available without restrictions and only data sets subject to IPR and confidentiality issues will be restricted. Where this is going to be the case, agreements will be made based on the individual data sets. Requests for the use of the data by externals will be approved by the project consortium.

2.4.D How the provenance of the data will be documented using the appropriate standards?

2.4.D a) Description

It is necessary to include information about entities, activities and people involved in producing data.

2.4.E Are data quality assurance processes described?

2.4.E a) Description

Describe what are your data quality assurance processes. How/when internal data quality assessments will be implemented?

The data quality can be ensured by different measures. These include validation of the sample, replication and comparison with results of similar studies and control of systematic distortion.

2.4.E b) Real example

Ex. 1 The quality of the dataset is guaranteed by the platform functioning.

Ex. 2 The data quality is ensured by different measures. These include validation of the sample, replication and comparison with results of similar studies and control of systematic distortion.

Ex. 3 Data quality assurance and control is central and the *raison d'être* of this project. About 80% of the efforts spent in our Thematic Centres is directed at data quality assurance.

3. Other research outputs

3.A Will there be other research outputs that may be generated or re-used throughout their project?

3.A a) Description

Explain what other research outputs have been generated in the execution of the project. They can be digital outputs such as software, workflows, protocols, models, etc. or physical outputs such as new materials, antibodies, reagents, samples, etc.

3.A a) Real example

Ex. 1 Where other research outputs are generated (only digital outputs are foreseen), the management of those products is shared with their creators

3.B Will there be other research outputs that may be generated or re-used throughout their project?

3.B a) Description

If you get other research outputs, explain how to make them findable, accessible, interoperable, and reusable. For more information, see 2. FAIR Data:

- Explain if other research outputs will be identified by a persistent identifier, will rich metadata provided to allow discovery, harvested, and indexed, what the metadata will be created and what disciplinary or general standards will be followed, will search keywords be provided in the metadata.
- Explain how these research output will be accessible from the repository, how open or restricted they will be, which license to use.
- Explain what data and metadata vocabularies, standards, formats, or methodologies will you follow to make your research outputs interoperables to allow these outputs exchange and re-use within and across disciplines.
- Explain how you will make these search outputs re-use. What documentation will you provide, what re-use license will you apply, etc.

4. Allocation of resources

4.A What will the costs be for making data or other research outputs FAIR in your project?

4.A a) Description

Estimate the costs for making your data FAIR.

4.A b) Real example

Ex. 1 There are no costs associated to the described mechanisms to make the database FAIR and long term preserved.

Ex. 2 The costs for depositing the dataset with the project, and subsequent resources required to make the dataset publicly available have been included within specific Work Packages within the project.

4.B How will these be covered?

4.B a) Description

State how you plan to cover the cost for making your data FAIR, including additional costs of archiving and preservation: direct and indirect costs related to storage, archiving, re-use, security, etc.

Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions).

4.B b) Real example

Ex. 1 <AcronymProject> is managed and supported by a team of experts and is free of charge.

Ex. 2 The cost of preserving the database will be assumed by the <Partner1>.

Ex. 3 (1) A dedicated hard disk drive will probably be allocated for the dataset. No costs are currently foreseen regarding its preservation.

(2) The cost will be covered at the local hosting institute in the context of the project.

(3) The cost will be covered at the local hosting institute as a part of the standard network system maintenance.

4.C Who will be responsible for data management in your project?

4.C a) Description

Explain the responsibilities for data management in your project.

4.C b) Real example

Ex. 1 The project coordinator has the ultimate responsibility for the data management in the project.

Ex. 2 Each partner has to respect the policies set out in this DMP. Datasets have to be created, managed and stored appropriately and in line with applicable legislation.

Ex. 3 The Project Coordinator has a particular responsibility to ensure that data shared through the website are easily available, but also that backups are performed and that proprietary data are secured.

Ex. 4 Quality control of these data is the responsibility of the relevant WP leader, supported by the Project Coordinator.

4.D How will long term preservation be ensured?

4.D a) Description

State how you plan regarding long-term preservation and who decides on what data will be kept and for how long.

4.D b) Real example

Ex. 1 Regarding the question of long-term data preservation, no specific arrangements has been done in the consortium yet. However, with a great degree of confidence, it can be confirmed that it is the project coordinator with the help of local <AcronymProject> resources who will play the major role in this task.

5. Data security

5.A What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?

5.A a) Description

Briefly describe the technical measures that will be implemented in the short to medium term to ensure data integrity (data backup), recoverability (prevention of data loss) and security (to prevent unauthorized access).

5.A b) Real example

Ex. 1 Data collected from the research group for the Project will be digitised and stored on the University's <StorageService> which is subject to regular back-up that is controlled by the University's IT personnel. The IT department performs operations by type: mission-critical (user data, virtual machines, scientific results, etc.) and static (scientific data sets, intermediate files, etc.). Content will be checked regularly to preserve its integrity, security, and durability. These procedures are designed, set and applied in order to fully comply with personal data as ruled by current national legislation and institutional regulations. Research team members will have an appropriate access level according to their role in the project.

5.B Will the data be safely stored in trusted repositories for long term preservation and curation?

5.B a) Description

Describe where the data will be safely stored in a trusted repository for long term preservation and curation. Also briefly describe the security setting of the chosen repository.

5.B b) Real example

Ex. 1 At the end of the project, all the final data will be deposited at the CORA.RDR data repository. The following list describes their security settings:

- *Versions: Data files are versioned. Records are not versioned. The uploaded data is archived as a Submission Information Package. Derivatives of data files are generated, but original content is never modified. Records can be retracted from public view; however, the data files and records are preserved.*
- *Replicas: All data files are stored Consorcio Madroño in Madrid. Data files are kept in multiple replicas in a distributed file system, which is backed up to tape on a nightly basis.*
- *Retention period: Items will be retained for the lifetime of the repository. The RDR has defined a lifetime for the repository of the next 10 years minimum.*
- *Functional preservation: The RDR makes no promises of usability and understandability of deposited objects over time.*
- *File preservation: Data files and metadata are backed up nightly and replicated into multiple copies in the online system.*
- *Fixity and authenticity: All data files are stored along with an MD5 checksum of the file content and the tabular file is stored with Universal Numerical Fingerprint (UNF).*
- *Files are regularly checked against their checksums to assure that file content remains constant.*
- *Succession plans: In case of closure of the repository, a guarantee has been made from RDR to migrate all content to suitable alternative institutional and/or subject-based repositories.*

6. Ethics

6.A Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?

6.A a) Description

Description of there is, or could there be, any ethics or legal issues that can have an impact on data sharing. These can be discussed in the context of the ethics review. If relevant include references to ethics deliverables and ethics chapter in the Description of the Action (DoA)

Specify if the informed consent for data sharing a long-term preservation is will be included in questionnaires dealing with personal data.

If your research activities involve children, patients, vulnerable populations, the use of human embryonic stem cells (hESCs) and human embryos (hEs), humans, Human cells or tissues, personal data, animals, Non-EU Countries, Environment, Health and safety, Artificial intelligence, Other ethics issues (man-machine interaction, develops in nanotechnology...), and Crosscutting issue: potential misuse of results (Activities that involve or generate materials, methods, technologies or knowledge that could be misused for unethical purposes) you must comply with ethical principles and relevant national, EU and international legislation.

6.A b) Real example

Ex. 1 All the activities carried out under the <AcronymProject> project comply with ethical principles and relevant national, EU and international legislation, for example the Charter of Fundamental Rights of the European Union and the European Convention on Human Rights. The tasks for <AcronymProject> only concern basic research activities and the project does not involve humans, animals or cells. Due to the fact that the main domain of the <AcronymProject> project activity is related to materials science with the focus on refractory materials, the risk of having ethics issues during the project is extremely limited. Either way, within the <AcronymProject> DoA Part A, the workpackage 8 is devoted to the ethics issues which sets out the 'ethics requirements' that the <AcronymProject> project must comply with. One deliverable will be provided: D8.1 NEC -Requirement No. 1. In the framework of D8.1, all beneficiaries and partner organisations must confirm that the ethical standards and guidelines of Horizon2020 will be rigorously applied, regardless of the country in which the research is carried out.

Ex 2. The transfer of data on human subjects to the <AcronymProject> repository is only considered when: informed consents, ethics approval and – when applicable - approval by local data protection authorities cover the purpose that the data are envisaged to be used within <AcronymProject> and allow transfer of individual or aggregated data to the <AcronymProject> repository. All data that are transferred to the <AcronymProject> repository shall be either pseudonymised or completely anonymized. The Data Owner/Data Provider is responsible for the anonymization or pseudonymation process and for ensuring that identifiable variables are not transferred to the <AcronymProject> repository. Directly identifiable variables include - but are not limited to - national ID number, name, phone number, ZIP-code, e-mail address, address, geographical coordinates (at a resolution that risks identification). One shall also be aware that a combination of just of few indirect identifying variables (such as birth data, gender, and zip-code) can be used to identify

a large portion of individuals on any dataset. In this context, the Data Owner/Data Provider shall only provide such variables at the lowest possible resolution that is necessary for analysis, e.g. district instead of zip-code; year of birth or age instead of birth date.

7. Other issues

7.A Do you, or will you, make use of other national/funder/sectorial/departamental procedures for data management?

7.A a) Description

Explain the national/funder/sectorial/departamental procedures for data management that you are using.

7.A b) Real example

Ex. 1 *As part of <University>'s commitment to ensuring FAIR and Open data, all research active staff (Postdoctoral fellows, PhD students) are expected to prepare DMPs for their own data, as per the University's Research Data Management Policy. The <University> data management policy defines research data as "the evidence that underpins the answer to the research question and can be used to validate findings regardless of its form." Thus, data covers quantitative and qualitative statements, raw data from measurements and derived data—either cleaned or extracted from a researcher's primary dataset or derived from an existing source.*

Ex. 2 *As well as European Commission policies on open data management, Project Partners must also adhere to their own institutional policies and procedures for data management:*

Ex. 3 *Each of the partners will follow their national and institutional procedures for data management, in addition to this <AcronymProject> DMP.*