

Plan de gestión de datos como requisito de las Agencias Públicas de Investigación

Un PGD es un plan de gestión de datos. La mayoría de las agencias financiadoras del área de ciencia, tecnología, innovación y desarrollo lo incluyen como requisito actualmente. Las agencias pretenden con esta política que los datos que se generan se preserven de forma segura durante la investigación financiada y que se almacenen a largo plazo, así como que se compartan de forma pública, siempre que este punto no entre en conflicto con necesidades de protección de datos. Un PGD requiere que el científico planifique la gestión de datos de investigación (GDI). En un PGD se explica qué datos se van a generar, en qué formatos se recogerán y cómo se pretende almacenarlos durante el desarrollo del proyecto, así como la forma en que se harán accesibles una vez finalizado el proyecto. La planificación de la GDI es la primera etapa para hacer que los datos sean localizables, accesibles, interoperativos y reutilizables (FAIR, por sus siglas en inglés).

Aunque se consideren distintas fuentes de financiación para un proyecto concreto, los requerimientos de las distintas agencias respecto al PGD probablemente diferirán poco entre sí. Existe un consenso general sobre qué aspectos deben describirse en un PGD:

- Características de los datos que se obtendrán, procesarán y/o generarán
- Identificación de las personas interesadas en los datos y qué papel jugarán en la obtención de los datos, procesamiento, etc., durante el proyecto
- Listado del manejo de los datos durante la ejecución del proyecto y las soluciones para su archivo a largo plazo
- Aspectos relevantes a la hora de compartir los datos, su propiedad y la privacidad necesaria

Estos aspectos se centran en mantener los datos de forma segura durante la ejecución del proyecto y, posiblemente, en su publicación y/o patente posterior a la ejecución. Este segundo aspecto de la accesibilidad refleja el interés de las agencias financieras en que los datos se ajusten a la ordenación FAIR.

Requisitos de los PGD: Redactar un PGD es obligatorio únicamente para algunas agencias, aunque se está convirtiendo en un requisito general. El Ministerio de Ciencia y Tecnología y el Ministerio de Sanidad ya lo han implementado. H2020 es uno de los planes de investigación que están pidiendo actualmente un PGD, y proporcionan un formulario para el mismo. Algunas Fundaciones Privadas, como la Fundación LaCaixa también piden un PGD. Para algunas agencias los PGD se presentan únicamente si el proyecto va a ser financiado, mientras que otras lo piden de forma previa, junto con la propuesta científica. En la mayoría de los casos, se requiere únicamente una sección breve sobre la gestión de los datos en la propuesta inicial.

Contenido del formulario para PGD: Algunas agencias financieras proporcionan su propio formulario de PGD, que debe ser cumplimentado por los investigadores. Los elementos que se incluyen son comparables al listado superior, basado en una gestión de datos según ordenamiento FAIR.

Dónde almacenar los datos a largo plazo: Aunque de forma general las agencias financieras permiten un amplio rango de opciones para almacenar los datos según las disciplinas científicas por las que son generados (sistemas de archivo seguro de datos, standard aceptados, sistema sostenible de archivo de datos), siempre aparece como requisito específico un índice, o archivo de metadatos (archivo que describe los datos generados; datos sobre los datos) que describe el origen de los grupos de datos generados y su contenido en los PGD.

Accesibilidad pública de los datos: Todas las agencias financieras requieren que los datos cumplan con las características FAIR y algunas de ellas van un paso más allá y piden que se haga de una forma abierta y universal (Open Access), en línea con los planes nacionales de Ciencia Abierta ('Open Science'). Sin embargo, todas las agencias financieras y de investigación también permiten el acceso restringido a los mismos siempre que sea necesario. Los principios que operan en estos casos son: las necesidades de privacidad y protección de datos, los intereses comerciales, la seguridad pública y los derechos de propiedad intelectual.

En general, respecto a qué datos hacer accesibles, las agencias financieras, las agencias de investigación e incluso las editoriales científicas requieren que todos los datos sean accesibles. Algunas de ellas únicamente exigen que sean públicos y accesibles aquellos datos que sean esenciales para la validación de los estudios realizados y/o que sean relevantes para su reutilización por otros investigadores o agentes sociales.

Plan de Gestión de Datos

Para cumplir con los estándar FAIR se requiere que los datos generados en los estudios de investigación sean localizables, accesibles, interoperables y reutilizables. Para ello, se deben incluir los siguientes apartados en el PGD, así como cualquier otra recomendación que se reciba por parte de las agencias financieras.

1. Características de los datos que se obtendrán, procesarán y/o generarán: describir en cada caso lo que atañe al proyecto de forma concisa y clara, en relación con la experimentación propuesta.

Se generarán datos provenientes de pacientes, donantes sanos, modelos y células animales de las siguientes técnicas:

Datos de imagen de distintas técnicas de microscopía, resonancia magnética, Ultrasonidos....

Datos de citometría de flujo, hemogramas,

Datos bioquímicos de ensayos electroforéticos, immunotransferencia, parámetros clínicos,....

Datos de perfiles metabolómicos, lipidómicos (HPLC-Spec)....

Datos de proteómica: espectrometría de masas, iTRAQ, SILAC,

Datos genómicos: RNAseq, secuenciación masiva, NGS,....

Datos epigenómicos: ATACseq, ChIPseq,

Los resultados directos de dichas técnicas (datos crudos; 'raw data') se harán accesibles como archivos originales generados por los sistemas y programas de los equipos empleados, así como formatos que sean utilizables por otros usuarios (por ejemplo: archivos FCS para citometría de flujo a partir de software BD Biosciences, LEI o LIF de Leica Microsystems, FASTQ y PDF para genómica, hojas Excel extendidas para proteómica, lipidómica y metabolómica, JPEG, TIFF, PNG para imágenes). Esto dependerá de los sistemas utilizados por cada equipo de trabajo.

2. Identificación de las personas interesadas en los datos y qué papel jugarán en la obtención de los datos, procesamiento, etc., durante el proyecto

El investigador principal (IP) se hará cargo de la supervisión de los datos generados por el grupo y del plan de gestión: GDI. En el GDI se indicará la persona que será asignada por el IP para la salvaguarda de la colección de datos generada y su distribución. Los datos se almacenarán en sistemas

computerizados/ digitales que tengan un acceso restringido mediante contraseña. Todos los pasos desde la generación de los datos hasta la obtención de los datos procesados se grabarán y almacenarán en estos sistemas. La persona responsable de cada uno de estos procesos dejará constancia en un registro de la actividad realizada de forma diaria en una base de datos que sea accesible para el resto del equipo de investigación y el IP.

3. Listado del manejo de los datos durante la ejecución del proyecto y las soluciones para su archivo a largo plazo

Los protocolos experimentales, el trabajo de investigación y los resultados obtenidos se registrarán de forma rutinaria y con todos los detalles necesarios en cuadernos de laboratorio, bien en formato papel, bien digital o ambos formatos. Las etiquetas incluirán el nombre del o de los investigadores que realizan la actividad, el investigador al cargo, el flujo de trabajo para cada uno de los protocolos realizados, los cálculos, las imágenes generadas y el texto.

El almacenaje y mantenimiento de los datos en formato digital se realizará por los expertos en tecnología informática designados por el IP, mediante el grabado en repositorios seguros internos o en repositorios externos seguros, encriptados y protegidos mediante contraseña que estén on-line (almacenaje en la nube), siguiendo el modelo establecido FAIR para los principios de la gestión de datos científicos.

4. Aspectos relevantes a la hora de compartir los datos, su propiedad y la privacidad necesaria

La interoperabilidad de los datos se garantizará mediante el uso del inglés (cuando se requiera) y de palabras clave y términos universales para la comunidad científica. Los datos de investigación generados se publicarán en la modalidad ‘Open Access’ siempre que sea posible y el acceso a los mismos se permitirá a través de licencias específicas.

Para su publicación todos los datos que resulten de las investigaciones realizadas se identificarán de forma inequívoca (DOI, handle, etc.) y se compartirán con la comunidad científica mediante el uso de un repositorio de datos institucional y/o temático. En el caso de nuestra institución, el IIS-IP, incluiremos los datos en el Repositorio Institucional de la Consejería de Sanidad de la Comunidad de Madrid y en cualquier otro que sea obligado por convocatoria o por las especificaciones o especificidades de los propios datos o del propio proyecto. Los servidores que den acceso a los datos almacenados estarán protegidos mediante contraseña. La Institución hará copias de seguridad regulares para evitar que se pierdan los datos. Todos los procedimientos que requieran autorización de comités éticos específicos habrán sido presentados, evaluados y aprobados por los mismos, así como por las autoridades pertinentes, según las leyes vigentes en España y la Unión Europea.

Una vez que se hayan publicado los datos en revistas especializadas, los datos procesados serán accesibles (si los términos de privacidad lo permiten) para la comunidad científica y el público general mediante los repositorios públicos que permiten el acceso a los manuscritos previos a la publicación y a los datos originales. Estos datos podrán ser accesibles también a través de repositorios públicos y/o acceso tras petición expresa al IP y la Institución.

Finalmente, se usará una estructura de metadatos para indexar los datos generados y hacerlos interoperables para la comunidad científica siguiendo el modelo de Dublín, de modo que se puedan realizar búsquedas masivas para su consulta a través de la red o de servidores especializados. Los datos publicados seguirán los estándares existentes para la identificación de los autores, ubicaciones y afiliaciones.

Casos particulares que se deban incluir:

Propiedad intelectual

Los métodos que se puedan desarrollar a partir de la investigación de los proyectos pueden tener repercusión en **técnicas de diagnóstico, modificaciones de tratamientos y protocolos existentes, medicamentos o vacunas.**

Los participantes en los proyectos deben hacer públicos sus resultados. Los resultados derivados de la investigación básica serán compartidos electrónicamente por la base de datos, cuyo enlace se publicarán en revistas indexadas, congresos, conferencias, informes, redes de investigaciones europeas y mundiales, etc. en que se participen los resultados.

Los datos se comunicarán inicialmente, de manera confidencial, a los miembros del proyecto, hasta su evaluación como potenciales productos o tecnologías por si pudieran ser patentables o debieran ser protegidos intelectualmente. Se considerarán patentables, en la medida de lo posible, procedimientos innovadores y nuevos productos, y se patentarán bajo acuerdo entre las instituciones que hayan participado en un proyecto conjunto. Los resultados no patentables se harán públicos lo más rápidamente posible.