Consejos y ejemplos para elaborar un Plan de Gestión de Datos (PGD) en el marco de la Convocatoria 2021 del Programa Estatal de Promoción del Talento y su Empleabilidad en I+D+I, Subprograma Estatal de Generación de Conocimiento, Proyectos de Investigación en Salud

ISPA

Instituto de Investigación Sanitaria del Principado de Asturias

	Elaboración
V.0	Comisión de Open Science
	Febrero 2020

ÍNDICE

¿Qué es un PGD?

El Plan de Gestión de Datos (PDG) o Management Data Plan (MDP) es un documento que aúna las declaraciones del proyecto de investigación con respecto a los datos que se van a utilizar en el mismo. El PGD debe especificar:

- Qué tipología de datos generará y recopilará el proyecto.
- Qué estándares se utilizarán.
- De qué forma se explotarán los datos y se compartirán, de forma accesible para su consulta, verificación y reutilización.
- Criterios de transparencia.
- Procedimientos de conservación y preservación de datos.

El PGD se creará al comienzo del proyecto de investigación y, según la evolución del proyecto, se completará o modificará en caso necesario a lo largo del ciclo de vida de la investigación.

Se trata, pues, de un documento descriptivo, ya que en él se reflejarán las distintas fases o ciclos de gestión de los conjuntos de datos utilizados, dinámico, es decir, que debe ser revisado en las distintas fases del proyecto (revisable), instrumental, ya que debe servir al control y verificación del propio proyecto así como a la asignación de fondos y recursos e histórico, teniendo en cuenta que debe indicar cómo se llevará a cabo la preservación de los datos una vez finalizado el proyecto así como los criterios de reutilización de los mismos.

¿Qué información debe consignarse en el PGD en el marco de esta convocatoria en la memoria de solicitud?

Esta convocatoria de 2021 de Proyectos de Investigación en Salud de la AES es la primera en la que se pide información sobre la gestión de datos en la memoria de solicitud. A continuación, se indican una descripción básica de la información a aportar en los distintos apartados así como ejemplos.

Describir la tipología y formato de los datos a recoger

En general, los datos a recoger pertenecerán a alguna de las siguientes categorías: observacionales, experimentales, simulados, operativos, derivados/compilados, del sector público o reutilizados, entre otras.

En su PGD deberá hacer una descripción detallada de cómo se generarán u obtendrán los datos, incluyendo también información sobre el formato de los archivos, el software que se utilizará y cómo se procesarán los datos.

Ejemplos:

-The experimental data from the study will be collected using a secure web-based system (MedSciNet), with external data streams being provided with a variety of software packages/formats including Microsoft Access (.mdb) or SQL, Microsoft Excel (.xlsx), DICOM for images (.dcm), uncompressed TIFF images (.tiff) and PDF for some research records (.pdf). Epigenetic Next Gen sequencing data will use BAM and BED files and pyrosequencing data will be collected using Microsoft Excel (.xlsx). RT-qPCR data analysis will be carried out using personal licenses for the software Sequence Detection Systems (SDS), ver. 2.4 (Applied Biosystems, USA) and GenEx qPCR data analysis software (MultiD Analysis, Sweden). R-Commander will typically be used for data analysis, but a wide variety of more specialist statistical software could also be used.

-Data will be published in .CSV format which is one of the most readable formats for information storage, supported by majority of software for numerical and data analysis.

-All raw and processed data will be published in proprietary file format (eg. PDF exports) and exported open format (eg. .csv, .txt or .md).

-Commonly used format: .txt, .doc(x), .xls(x), .jpg/.jpeg/.png, .tiff, .pdf, .csv

-Data will be documented/described including all methodology. Data generated during the project will be accompanied by standardised, structured metadata record explaining the purpose, origin, creator(s), access conditions and terms of use of the data. Metadata of a minimum Dublin Core standard will be produced

-Open source standard formats will be used wherever possible. Otherwise standards in the field will be used. Records data on up to 70,000 subjects:

- *SQL databases for medical records data*
- *FASTQ for mRNA or DNA sequencing data*
- *GEN files for genotype data*
- *Digital pathology images will be in Hamamatsu format NDPI, and Aperio SVS format,Ventana and other vendors TIF*
- *Quantitative CSV, text file of mRNA expression for Nanostring*

Generar en el marco del proyecto el procedimiento previsto para acceso a los datos (quién, cómo y cuándo podrá acceder a ellos)

En el PGD del proyecto se deberá describir cómo se compartirán los datos, incluyendo el procedimiento de acceso, los periodos de embargo (si los hay) y definir si el acceso será totalmente abierto o restringido. Además, en el caso de que el acceso a algunos datos esté limitado, deberá justificarse el porqué: razones éticas, protección de datos personales, implicación de derechos de propiedad intelectual y/o industrial, intereses comerciales, etc.

Ejemplos:

-Once processing, quality control, organisation, analysis and publication are complete, the data will be made accessible by deposition in open access repositories. These data will be anonymized

so as not to have any potential correlation and identification of the ethical issues with their publication and dissemination.
The data cannot be publicly shared because it contains potentially identifying information of human subjects.
The data we propose to collect will be suitable for sharing with national and international scientists. Once we have completed the aims and objectives of this proposal, we would open the dataset up for collaborators to use within the ethical and research governance constraints of the study (any datasets released to collaborators will always be anonymised).
Data cannot be released until the patents related to this research are issued.

-Data and software underpinning research articles will be available to other researchers at the time of preprint submission prior to publication in journals, providing this is consistent with:

- *any ethics approvals and consents that cover the data (meta data will only be used for electronic health records data in line with ethics and privacy guidelines)*
- *reasonable limitations required for the appropriate management and exploitation of IP.*

-the final data has no sensitive data, "all people can access the final result"

Titularidad de los datos y repositorio en que se prevé realizar su depósito, difusión y preservación

Titularidad:

Sebe indicar sobre qué institución o investigador recae la titularidad de los datos, que será responsable de su custodia y difusión.

Ejemplos:

-Experienced researcher/ PI is responsible for data management.

-The project PI will be responsible for the governance of the research data access during the project period.

Depósito, difusión y preservación:

En la memoria se debe hacer constar el modo de publicación: como dataset exento depositado en un repositorio, como información suplementaria de un artículo, o como data paper. Los datos de investigación no solo se difunden mediante repositorios, se difunden mediante enlaces desde el artículo de revista al repositorio donde se ubica el dataset. Pero en bastantes casos también se incluyen como material complementario del artículo en la revista. Y también han surgido data journals o revistas de datos, especializadas en diseminar estos materiales, que se publican en este caso en forma de artículos de datos o data papers. La difusión de un dataset a través de revistas es una vía complementaria a su depósito en repositorios, para los casos en que su revisión por pares y publicación tiene un interés intrínseco adicional.

Ejemplos:

-The data from this study will be stored on the … servers. A formal backup regime will be followed, with the procedure documented. A complete backup of the active data drives will be made each night and backup tapes will be taken offsite on a weekly basis. These are collected on a Grandfather/Father round robin system with a monthly backup being held in a fireproof safe. Data will be stored using standard software, and generic procedures will be developed such that data from all studies can be readily accessed by authorised staff.

-Data will only be shared as supplementary material via publication.

-the data will be found on Zenodo, there is no restriction on the access on the data, the license of the reulted data is CreativeCommons Attribution 4.0 International

-The processed anonymized and codified data and their metadata will be shared in an open research data repository without any restriction (Creative Commons BY).

- All data generated from this action along with suitable metadata information will be stored in an open access data repository eg. Zenodo and also University of…….. data repository. A unique digital object identifier (DOI) will be given to each data uploaded to ease the identification mechanism.

-Zenodo, an open access repository, will be used for making data FAIR.

Procedimiento previsto para garantizar los requisitos éticos o legales específicos de aplicación (ej. privacidad de los datos y su reglamentación; datos protegidos o protegibles por propiedad intelectual o industrial, etc.) que condicionen su disponibilidad, uso y/o reutilización

Existen numerosas razones por las que sus datos pueden ser sensibles, como las cuestiones éticas que rodean a los datos personales o biométricos, consideraciones legales como los derechos de propiedad intelectual, intereses comerciales o patentes, cuestiones medioambientales como la localización de especies en peligro de extinción o aspectos relacionados con la seguridad para organizaciones o países. Cualquiera de ellos debe ser identificado y abordado en su PGD, a pesar de lo cual sus datos pueden compartirse si se toman las medidas necesarias para tratarlos de forma responsable y adecuada.

Entre las estrategias para tratar los datos sensibles, que deben consignarse en su PGD, se encuentran: obtener la aprobación de los comités de ética correspondientes, cumplir con la legislación de protección de datos, generar y recopilar los consentimientos informados, anonimización, acceso controlado y almacenamiento seguro.

Ejemplos:

-The project was analysed by the Ethical Committee for Health (Comissão de Ética para a Saúde, CES) and received a favourable opinion.

-It is not necessary that patients sign an informed consent statement because this study is observational, not interventive.

-Data collection sheets will not contain any personal identifiable information.

-No clinical records will be taken off site. All data recorded on data collection sheets will be completely anonymised and it will not be possible to trace this information back to the patients.

-All collected data will be completely anonymised

-A list of the identification codes of patients with acute myeloid leukemia will be stored in an excel file with a password, together with the rawtext data, in a specific folder (Folder B) in a computer with an encrypted disk

-The files and their contents will be permanently deleted at the end of the work, and the collected data (raw text data) will be anonymizedand codified.

Derechos de autor

-An agreement between the Principal Investigator and the other project collaborators regarding Intellectual Property Rights (IPR) ownershipissues will be written and signed

-When the study is published the authors will be the owner and have the copyright

ISPA

Instituto de Investigación Sanitaria
del Principado de Asturias