

Plan de Gestión de Datos

Proyecto PID2020-114153GB-I00

Título	Adaptación a la variación ambiental predecible e impredecible en rotíferos: integración de estrategias y escalas temporales
Acrónimo	UNPREDSCALES
Responsables	Eduardo M. García Roger (IP1) María José Carmona Navarro (IP2)
Fecha	19/11/2025
Contacto	eduardo.garcia@uv.es maria.j.carmona@uv.es

Versión	Fecha	Descripción	Autores
1.0	19/11/2025	Documento inicial	Eduardo M. García Roger Maria J. Carmona Navarro

Contenido del documento

Información del documento	1
I. Resumen de Datos – Datos Administrativos.....	3
A. Propósito	3
B. Relación con los objetivos del proyecto	3
C. Tipología, origen y formato de datos	3
D. Re-Utilización	3
E. Tamaño de los datos	3
II. Datos FAIR	7
A. Localización de los datos [Findable Data]	5
B. Accesibilidad de los datos [Accesible Data].....	5
C. Interoperabilidad de los datos [Interoperable data]	6
D. Reutilización de los datos [Re-usable data].....	6
III. Disposición de recursos	7
IV. Seguridad	7
V. Aspectos éticos	7

I. Resumen de Datos

A. Propósito

Este informe describe el Plan de Gestión de Datos (PGD) del proyecto UNPREDSKALES (PID2020-114153GB-I00), financiado por el Ministerio de Ciencia, Innovación y Universidades del Gobierno de España a través de la Agencia Estatal de Investigación como parte del Programa de Proyectos I+D+i 2020 para la Generación de Conocimiento. Los coordinadores de este PGD son los investigadores principales de dicho proyecto, el Dr. Eduardo M. García Roger y la Dra. María José Carmona Navarro, profesores del Departamento Microbiología y Ecología de la Universidad de Valencia (UV), e investigadores del Instituto Cavanilles de Biodiversidad y Biología Evolutiva (ICBiBE-UV).

El propósito de este PGD es proporcionar una descripción detallada de los procedimientos y protocolos para la gestión de los conjuntos de datos generados durante el desarrollo del proyecto. Este PGD describe los principales principios de gestión de datos en términos de estándares y metadatos, compartición, archivo, preservación y seguridad.

Se trata de un documento vivo que ha sido desarrollado durante el período de ejecución del proyecto y que continuará revisándose conforme avancen la publicación y los resultados de los distintos objetivos. Se espera, además, que los datos generados puedan reutilizarse, por lo que el documento se actualizará de manera periódica más allá del periodo de ejecución del proyecto. Este PGD y sus futuras actualizaciones se alojarán en el repositorio institucional de la UV, RODERIC, bajo el nombre UNPREDSKALES_PGD_V1.0.pdf (véase la sección 2.1 para las convenciones de nomenclatura).

B. Relación con los objetivos del proyecto

El proyecto UNPREDSKALES ha generado varios conjuntos de datos de distintos tipos, tanto cuantitativos como cualitativos. El PGD servirá para apoyar los objetivos científicos del proyecto y difundir sus resultados. Esto incluye la gestión de tres categorías principales de datos:

1. **Objetivos de investigación.** Los conjuntos de datos asociados a esta categoría permitirán a cualquier usuario potencial replicar los principales resultados científicos del proyecto. Esto incluye datos brutos de datos ambientales y experimentos, simulaciones por ordenador, y códigos para producir, descargar y analizar datos.
2. **Actividades de difusión para audiencias expertas.** Los conjuntos de datos asociados a esta categoría consisten en documentos que resumen los principales resultados científicos del proyecto. Esto incluye artículos científicos publicados en acceso abierto, preprints, informes y presentaciones en congresos.
3. **Actividades de comunicación para audiencias no expertas.** Los conjuntos de datos asociados a esta categoría consisten en documentos destinados a fines educativos y divulgativos.

C. Tipología, origen y formato de datos

El proyecto UNPREDSCALES investigó los rasgos relacionados con la salida de la diapausa en rotíferos y su asociación con la predecibilidad de los hábitats que ocupan. Para ello, Se utilizaron dos enfoques experimentales complementarios cuyo objetivo era obtener datos sobre la fracción y el tiempo de eclosión de los huevos de diapausa. El primer enfoque generó datos mediante el análisis correlacional entre rasgos fenotípicos de clones de rotíferos procedentes de poblaciones naturales, su genotipo y parámetros hidrológicos de sus localidades de origen. El segundo enfoque produjo datos a través de experimentos de evolución en laboratorio, en los que clones de rotíferos fueron sometidos a ciclos de selección bajo diferentes regímenes de predecibilidad ambiental. En ambos casos, la caracterización de los rasgos se realizó mediante bioensayos de laboratorio. Estas dos aproximaciones han generado conjuntos de datos estructurados y no estructurados.

Los datos estructurados incluyen principalmente archivos numéricos —que, por simplicidad y operatividad, se denominarán así aunque también contengan variables categóricas como niveles de factores experimentales— y abarcan tanto los registros en bruto como los datos procesados generados por la investigación. Los datos brutos proceden de mediciones y registros experimentales, y recopilación de variables ambientales tomadas en el campo. Estos datos se organizaron siempre en forma de matrices o *dataframes* en hojas de cálculo. Por su naturaleza simbólica, compleja y dependiente de formatos especializados, los datos genómicos de secuenciación no pueden integrarse directamente en los *dataframes* numéricos utilizados en los análisis experimentales y deben considerarse una tipología diferenciada. A diferencia de los datos numéricos que se organizan en matrices interpretables, los datos brutos de secuenciación consisten en cadenas de bases (A, C, G, T) y archivos derivados que requieren herramientas bioinformáticas específicas, manejo de metadatos técnicos y un procesamiento computacional previo para transformarlos en información utilizable. Los datos elaborados, por su parte, surgieron del procesamiento computacional de los datos brutos, tanto numéricos como genómicos, para depuración, estandarización, cálculo de métricas, organización y codificación. El resultado final se estructuró igualmente en *dataframes* con el fin de facilitar su análisis estadístico.

El proyecto también incorporó datos no estructurados, principalmente imágenes satelitales provenientes de los programas LANDSAT y SENTINEL, descargadas de repositorios del USGS (Servicio Geológico de Estados Unidos) y de la ESA (Agencia Espacial Europea). Estas imágenes contienen valores radiométricos de reflectancia organizados en matrices de píxeles y se almacenan en formatos raster que permiten gestionar múltiples bandas y metadatos geoespaciales. Tras su descarga, las imágenes se procesaron computacionalmente para caracterizar parámetros hidrológicos y estimar la predecibilidad de los hábitats de las poblaciones de rotíferos analizadas.

Además de los datos destinados al análisis científico, el proyecto generó un conjunto adicional de información de tipo documental o textual combinado. Este incluye tablas, figuras y representaciones gráficas correspondientes a los resultados principales, así como los materiales suplementarios, asociados a los distintos objetivos de investigación.

El proyecto UNPREDSCALEs ha generado datos en diferentes formatos, que se presentan en la tabla siguiente.

Tipo de datos	Descripción	Formato	Justificación
Datos numéricos	Datos experimentales y ambientales, brutos y elaborados	.csv	- Eficiencia en el almacenamiento de datos. Se trata de un formato ligero y basado en texto plano que permite manejar grandes volúmenes de información sin requerir recursos adicionales. - Amplia compatibilidad con distintos programas, incluyendo hojas de cálculo, software estadístico y herramientas de análisis, tanto libre como comercial. - Alta interoperabilidad entre plataformas. - Accesibilidad, estabilidad, reproducibilidad a largo plazo: Formato abierto y no dependiente de aplicaciones propietarias
Datos genómicos	Datos de secuenciación, brutos y elaborados	FASTQ, SAM/BAM, VCF	- FASTQ: Almacenamiento de datos crudos de secuenciación, registrando tanto secuencias como la calidad de cada base, permitiendo evaluar la fiabilidad de las lecturas. - SAM/BAM: Organización y posicionamiento de lecturas tras el alineamiento a genomas de referencia. - VCF: Identificación y almacenamiento de variantes. Formato estandarizado que permite registrar tipo, posición y frecuencia de cada cambio genómico.
Códigos	Líneas de código (<i>scripts</i>) para manejo, análisis de datos y descarga automatizada de datos a partir de plataformas API.	.R, .py, .txt, bash	- R, como lenguaje de programación y análisis de datos de acceso abierto, permite la reutilización transparente del código y su modificación para analizar nuevos conjuntos de datos. - Compatibilidad con formato .txt - Invocación de ficheros .csv de almacenamiento de datos numéricos. Ejecución inmediata y reproducible. - Python y bash: Desarrollo <i>ad hoc</i> de <i>scripts</i> para manejo y descarga automatizada de datos
Imágenes	Datos organizados en matrices de píxeles (raster)	JPG, TIFF	- JPG: Compresión eficiente de grandes volúmenes de datos. - TIFF: Fidelidad y soporte de múltiples bandas. Inclusión de metadatos geoespaciales.
Datos comprimidos	Almacenamiento de datos	TAR.GZ	- Ahorro de espacio de almacenamiento - Empaquetamiento de archivos con contenido similar y/o complementario.
Documentales	Artículos, preprints, informes, presentaciones	PDF, .docx, .pptx	- Archivos documentales o textuales compuestos. PDF preserva calidad y formato en cualquier dispositivo. Integra distintos elementos en un único documento facilitando su lectura o visualización

D. Re-Utilización

El proyecto UNPREDSCALES contempló desde su inicio la posibilidad de reutilización de los datos generados, ya que pueden resultar relevantes para establecer un marco de referencia, realizar comparaciones y/o complementar los datos producidos por otros grupos de investigación. Para posibilitar la reutilización de los datos generados, se ha promovido la publicación en abierto de los resultados de investigación y el almacenamiento de datos en repositorios de acceso público con licencias de reconocimiento de la autoría, que garantizan la calidad de los datos y aseguran la preservación de estos. No se anticipan restricciones legales de acceso o éticas, como se indicó en la solicitud del proyecto. Además, se ha trabajado con formatos compatibles y estandarizados de almacenamiento de datos para facilitar su integración.

Los datos de rasgos asociados a la salida de la diapausa en rotíferos podrían resultar útiles para otros investigadores interesados en replicar los estudios o realizar análisis comparativos en otros sistemas de estudio, considerando otras localidades u otros organismos, pero también pueden ser de interés para instituciones y responsables de políticas públicas que pueden beneficiarse de los análisis de las series temporales de datos hidrológicos recogidos para el reconocimiento de tendencias, etc. En cuanto a los datos genómicos pueden servir como insumos en nuevas investigaciones o aplicaciones a mayor escala, incorporándose a grandes bases de datos de biodiversidad, y/o colecciones anotadas de secuencias de DNA como GenBank (NCBI - Estados Unidos) o EMBL-EBI (en Reino Unido).

E. Tamaño de los datos

La cantidad total estimada de datos en el proyecto UNPREDSCALES es de aproximadamente 170 GB. En las siguientes tres tablas resumimos, los diferentes tipos de conjuntos de datos generados en el proyecto, vinculados a los objetivos de trabajo del proyecto según se han presentado en el informe final.

Objetivo 1. Estudiar la expresión conjunta de la plasticidad fenotípica y las estrategias de minimización de riesgos (<i>bet hedging</i>) en rasgos relacionados con el momento de eclosión de los huevos diapáusicos en rotíferos			
Tipo de conjunto de datos	Origen	Formato	Tamaño
Datos brutos de rasgos de diapausa	Experimental	.csv	2 MB
Datos elaborados de rasgos de diapausa	Computacional	.csv	2 MB
Imágenes satelitales de LANDSAT y SENTINEL	Bases de datos públicas	TAR.GZ (incluye TIFF, JPG, TXT)	10 GB
Datos hidrológicos elaborados y métricas de predecibilidad	Computacional	.csv	2 MB
<i>Scripts</i> para manejo de datos, descarga y análisis estadístico (varios archivos)	Computacional	.R, .py, bash	< 1 MB
Archivos documentales	Experimental, Computacional	.docx, PDF	40 MB

Objetivo 2. Estudiar experimentalmente la respuesta evolutiva de las estrategias de diversificación de riesgos (*bet hedging*) entre estaciones y dentro de una misma estación en el momento de la eclosión de los huevos diapáusicos en los rotíferos.

Tipo de conjunto de datos	Origen	Formato	Tamaño
Datos brutos de rasgos de diapausa	Experimental	.csv	2 MB
Datos elaborados de rasgos de diapausa	Computacional	.csv	2 MB
Datos genómicos brutos de secuenciación masiva	Computacional	TAR.GZ (incluye FASTQ)	70 GB
Datos genómicos elaborados (varios archivos)	Computacional	TAR.GZ (SAM/BAM, VCF), .csv	40 GB
Scripts para manejo de datos y análisis estadístico (varios archivos)	Computacional	.R	< 1 MB
Archivos documentales	Experimental, Computacional	.doc, PDF	40 MB

Objetivo 3. Evaluar cómo las estrategias de minimización de riesgos en el momento de la producción de huevos diapáusicos contribuyen a la adaptación y éxito colonizador de los rotíferos en ambientes temporalmente variables.

Tipo de conjunto de datos	Origen	Formato	Tamaño
Datos brutos de rasgos de producción de huevos de diapausa	Experimental	.csv	2 MB
Datos genómicos brutos de secuenciación masiva	Computacional	TAR.GZ (incluye FASTQ)	30 GB
Datos genómicos elaborados (varios archivos)	Computacional	TAR.GZ (SAM/BAM, VCF), .csv	10 GB
Scripts para manejo de datos, descarga y análisis estadístico (varios archivos)	Computacional	.R, .py, bash	< 1 MB
Archivos documentales	Experimental, Computacional	.doc, PDF	20 MB

II. Datos FAIR

A. Localización de los datos [Findable Data]

Para garantizar la localización de los datos, se han aplicado diversas medidas alineadas con los principios FAIR. Los datos brutos y elaborados generados durante el proyecto UNPREDSKALES se han depositado en RODERIC, el repositorio institucional de la UV. Cada objeto en RODERIC

recibe un identificador persistente tipo Handle y se acompaña de metadatos descriptivos, como título, autores, palabras clave, fecha de creación y descripción del contenido. Esto permite una búsqueda precisa y la recuperación inequívoca de los datos dentro del repositorio y mediante motores de búsqueda académicos.

A efectos de publicación en revistas científicas indexadas, atendiendo a los compromisos de transparencia y reproducibilidad, se han publicado los datos asociados a las publicaciones en Zenodo, que genera automáticamente metadatos completos (título, autores, descripción, fecha, versión, licencia, palabras clave) y asigna un identificador DOI, garantizando la citación estable y la localización permanente de los conjuntos de datos. Finalmente, todas las publicaciones derivadas del proyecto cuentan con DOI, lo que facilita la vinculación entre artículos y datos, reforzando su accesibilidad y rastreabilidad. Asimismo, los datos genómicos se han publicado en las bases de datos de NCBI con enlaces de acceso para propietarios y revisores, y, una vez que se publiquen los resultados, se pondrán a disposición del resto de la comunidad científica.

Por su parte, los datos imágenes satelitales provenientes de los programas LANDSAT y SENTINEL, pueden descargarse de los repositorios originales del USGS (Servicio Geológico de Estados Unidos) y de la ESA (Agencia Espacial Europea). El acceso público a los datos de ambas agencias requiere un registro previo de usuario, pero es completamente gratuito.

B. Accesibilidad de los datos [Accesible Data]

El programa Horizon 2020 (H2020) de la Unión Europea promueve la investigación e innovación con un enfoque en la transparencia, accesibilidad y reutilización de los resultados científicos. Dentro de sus directrices, establece que tanto las publicaciones como los datos generados deben ser gestionados de manera que permitan un acceso abierto y seguro, garantizando su preservación y disponibilidad a largo plazo. En este marco, se han aplicado las licencias de la organización sin ánimo de lucro *Creative Commons* tanto a las publicaciones de datos como a los artículos derivados del proyecto con el fin de garantizar el acceso abierto amplio y conforme a las directrices de la UE y de las plataformas de depósito.

En particular, se recomienda el uso de *Creative Commons 4.0*, preferentemente la licencia CC BY-NC-SA (Creative Commons Atribución-No Comercial-Compartir Igual), de acuerdo con las prácticas habituales en repositorios de investigación y alineada con las directrices de OpenAIRE, el proyecto europeo de apoyo a la ciencia abierta financiado por la Comisión Europea y el European Research Council. En relación con los datos, esta elección se justifica porque repositorios como Zenodo y RODERIC permiten y promueven su difusión bajo esta licencia, facilitando así su reutilización y asegurando el reconocimiento de la autoría. En cuanto a las publicaciones científicas, el uso de esta misma licencia responde a los convenios establecidos entre las editoriales y la UV, que contemplan la publicación en acceso abierto completo (*Gold Open Access*) o híbrido (*Hybrid Open Access*) bajo condiciones CC BY. Cuando no existe convenio con la editorial de la publicación, se ha optado por una estrategia de acceso abierto verde (*Green Open Access*), un modelo de acceso abierto a publicaciones científicas en el que los autores depositan una versión de su trabajo (normalmente el manuscrito aceptado tras la revisión por pares, pero antes de la maquetación final de la editorial) en un repositorio institucional o temático, donde puede ser consultado libremente por cualquier persona, generalmente después de un periodo de embargo establecido por la editorial.

En conjunto, estas estrategias aseguran la máxima visibilidad y accesibilidad de los resultados del proyecto, al tiempo que cumplen con las directrices comunitarias y requisitos normativos vigentes.

C. Interoperabilidad de los datos [Interoperable data]

En la tabla siguiente se muestran ejemplos de programas para abrir y/o utilizar los datos generados por el proyecto UNPREDSKALES en diferentes sistemas operativos (Ubuntu, macOS y Windows) según el formato de almacenamiento de los datos. Estos ejemplos no son los únicos y el usuario potencial podría utilizar otros programas y/o plataformas estándar, ya que todos los formatos de archivos de datos utilizados en el proyecto son compatibles con distintos programas —incluyendo hojas de cálculo, software estadístico y diversas herramientas de análisis, tanto libres como comerciales—, lo que facilita su integración en múltiples entornos de trabajo. Además, al tratarse de formatos abiertos y no dependientes de aplicaciones propietarias, garantizan una alta interoperabilidad entre plataformas y aseguran un acceso absoluto a la información. Al menos una de las opciones para cada formato es de software libre (Ubuntu).

Formato	Ejemplos of software (Sistema operativo) para manejo
.csv	LibreOffice (Ubuntu, MacOS, Windows), Numbers (MacOS), Office (Windows)
JPEG, PNG, TIFF	gThumb (Ubuntu), Preview (MacOS), Adobe Illustrator (Windows)
PDF	Adobe Reader (Ubuntu, Windows), Preview (MacOS)
.py	Python sources releases (Ubuntu, MacOS, Windows) in www.python.org
.R	R sources releases (Ubuntu, MacOS, Windows) in https://cran.r-project.org/
TAR.GZ	TAR.GZ package (Ubuntu), tar utility (MacOS), 7-Zip (Windows)

D. Reutilización de los datos [Re-usable data]

Los conjuntos de datos estarán disponibles para su reutilización y se almacenarán en RODERIC y Zenodo sin ningún coste. Si los conjuntos de datos se actualizan, los coordinadores de este PGD serán responsable de gestionar las diferentes versiones, garantizando que la versión más reciente esté siempre disponible.

Las políticas adoptadas por RODERIC y Zenodo en relación con licencias, periodos de disponibilidad y calidad son las siguientes:

1. **Licencia.** Todo el material publicado en el repositorio RODERIC, así como en Zenodo, incorpora licencias de la organización sin ánimo de lucro Creative Commons. En particular, se recomienda el uso de la licencia CC BY-NC-SA, que se ha utilizado en el proyecto UNPREDSKALES y también en otros proyectos anteriores del grupo de investigación. Estas licencias permiten la reutilización de los datos tras la finalización del proyecto y su uso por parte de terceros.

2. **Disponibilidad.** Los datos permanecerán reutilizables una vez finalizado el proyecto, sin limitaciones temporales ni restricciones de acceso, salvo que se establezca expresamente un embargo o restricciones específicas. En el caso de Zenodo, los datos depositados reciben un DOI permanente que garantiza su accesibilidad y citabilidad a largo plazo.
3. **Calidad, curación y mantenimiento.** La calidad de los conjuntos de datos está garantizada por los sistemas de operación de RODERIC, que realizan copias de seguridad rutinarias y verificaciones de la integridad del material alojado. Por su parte, Zenodo aplica procedimientos estandarizados de preservación digital, que incluyen controles de integridad, mantenimiento del DOI, almacenamiento redundante y monitorización continua. Además, ofrece un sistema transparente de gestión de versiones, permitiendo el depósito de nuevas versiones sin eliminar las anteriores, lo que facilita la trazabilidad y la correcta citación de cada actualización.

III. Disposición de recursos

El alojamiento de los conjuntos de datos en RODERIC y Zenodo no implica costes, ni a corto ni a largo plazo. Se considera la posibilidad de alojar versiones de los preprints producidos a BioarXiv (un repositorio educativo privado sin fines de lucro, propiedad y gestionado por la Universidad de Cornell). Para la versión en BioarXiv se utilizará el mismo título de preprint que el empleado para la versión alojada en RODERIC. Con esto se asegurará y reforzará el acceso abierto verde de acuerdo con las directrices de H2020.

IV. Seguridad

La seguridad y preservación de los datos generados en el proyecto UNPREDSKALES se asegura a distintos niveles.

En primer lugar, Los conjuntos de datos generados por el proyecto se alojan en repositorios de reconocido prestigio como RODERIC, Zenodo y BioProjects del NCBI (para datos genómicos). Estos repositorios implementan avanzados programas de autopreservación, que incluyen copias de seguridad periódicas, verificación de integridad de los datos y control de versiones, asegurando que los conjuntos de datos se mantengan disponibles, completos y consistentes a lo largo del tiempo. Además, en el caso de RODERIC los datos reciben el mismo tratamiento de seguridad y gestión que el resto de los documentos alojados, garantizando su preservación dentro de estándares internacionales de repositorios académicos.

Los archivos originales de los autores, incluyendo datos brutos, procesados y documentación asociada, se conservan en los ordenadores personales de los coautores de cada objetivo de investigación del proyecto. Esta estrategia se complementa con el uso de OneDrive (Microsoft 365) con licencia institucional de la UV, que permite sincronización automática, redundancia en la nube y control de acceso mediante credenciales corporativas, asegurando así la disponibilidad y recuperación ante posibles fallos locales o pérdidas de datos.

El grupo de investigación de Ecología Evolutiva dispone de un espacio de almacenamiento en disco virtual (disco.uv.es) mantenido por el Servicio de Informático de la Universidad de Valencia (SIUV), gestionado por el propio grupo de investigación, que ofrece un entorno seguro y controlado para el almacenamiento de datos en fase de análisis y uso activo. El acceso está protegido mediante usuario y contraseña, y los permisos de edición se restringen exclusivamente a los miembros del equipo investigador, garantizando la integridad de los datos durante su manipulación y evitando modificaciones no autorizadas.

Como medida adicional, los datos se almacenan también en memorias físicas externas de propiedad de los autores de los trabajos. Estas unidades permiten mantener copias locales de seguridad que pueden ser utilizadas para recuperación rápida en caso de fallo de sistemas informáticos o incidentes imprevistos, reforzando la estrategia de preservación redundante en distintos niveles.

Finalmente, se hace uso del espacio de almacenamiento NAS del SIUV, que proporciona un entorno centralizado de almacenamiento en red con alta disponibilidad. Este sistema permite gestión de acceso controlada, manejo de grandes volúmenes de datos, especialmente para los datos genómicos derivados de la investigación. Se usa también para guardar copias de seguridad adicionales de los datos, asegurando que la información generada durante el proyecto esté protegida, preservada y disponible a largo plazo para todo el equipo de investigación.

V. Aspectos éticos

No aplica.