



GUÍA ORIENTATIVA PARA LA DOCUMENTACIÓN Y ORGANIZACIÓN DE CONJUNTOS DE DATOS

1. Introducción

La Universidad de Santander, a través de la Vicerrectoría de Investigaciones, promueve prácticas de ciencia abierta, gestión responsable de datos, documentación, trazabilidad y reutilización del conocimiento científico. En la convocatoria para el diseño y publicación de conjuntos de datos con potencial de explotación en Inteligencia Artificial se establecen como resultados el Conjunto de Datos en Formato Abierto, la Documentación del Conjunto de Datos, el Documento Técnico y el Documento con estudio prospectivo.

Esta guía orienta a los investigadores sobre qué se entiende por documentación del conjunto de datos y cómo organizar sus productos y evidencias en un repositorio institucional de GitHub. La propuesta busca que los proyectos utilicen una estructura común y que, progresivamente, la Universidad pueda consolidar un Catálogo Institucional de Conjuntos de Datos de Investigación.

GitHub se plantea principalmente como espacio de documentación, organización, trazabilidad y descubrimiento. El conjunto de datos no se almacenará directamente en este repositorio; deberá publicarse en el repositorio de datos que corresponda según su naturaleza, tamaño y condiciones de acceso.

Con el fin de precisar la información de la guía, a continuación, se abordan conceptos y diferencias entre conceptos importantes para comprender la esta guía orientativa:

2. Marco de conceptos

Concepto	Definición orientativa
Conjunto de datos o Dataset	Datos generados, recopilados, transformados, curados o anotados por el proyecto (encuestas, registros, imágenes, textos, sensores, datos geoespaciales, entre otros).
Documentación del conjunto de datos	Información que permite comprender el contenido y estructura del dataset: metadatos, diccionario de datos y muestras documentadas.



Diccionario de datos	Descripción de las variables, campos o dimensiones (nombre, significado, tipo, unidad, valores permitidos, codificación, entre otros).
Metadatos	Información que describe el dataset como recurso (título, autores, proyecto, palabras clave, cobertura, versión, licencia, acceso, repositorio, entre otros).
Documento Técnico	Documento que describe el objetivo, justificación, metodología de recolección y procesos de curación y anotación de los datos y metadatos.
Estudio prospectivo	Análisis del potencial futuro de explotación del dataset, incluyendo modelos de IA y casos de uso en diferentes áreas del conocimiento.
Repositorio	Espacio tecnológico para almacenar, preservar, organizar o facilitar el acceso a recursos digitales.
Catálogo Institucional	Organización de los conjuntos de datos de diferentes proyectos para facilitar su identificación, consulta, divulgación y reutilización.

3. Diferencia entre los principales productos

Producto	Pregunta que responde	Ejemplo
Dataset	¿Cuáles son los datos?	Archivo CSV con registros de una encuesta.
Documentación del dataset	¿Qué contiene y cómo se interpreta?	Metadatos, diccionario y muestras.
Documento Técnico	¿Cómo se generó y preparó?	Metodología de recolección, curación y anotación.
Estudio prospectivo	¿Qué potencial futuro tiene?	Modelos de IA, casos de uso y oportunidades.
Repositorio de datos	¿Dónde está publicado el dataset?	Repositorio institucional, Zenodo, Dataverse u otro.
Catálogo institucional	¿Qué datasets tiene la Universidad?	Registro organizado de los datasets de los proyectos.



Para la Vicerrectoría de investigaciones, la organización homogénea de los conjuntos de datos, permite que los productos de los proyectos trasciendan el cumplimiento de una convocatoria y puedan convertirse en recursos reutilizables para la comunidad académica y científica.

Desde la perspectiva de ciencia abierta, esta organización favorece la transparencia, la accesibilidad y la reutilización responsable de los datos. Desde la perspectiva de divulgación e impacto, facilita que otros investigadores encuentren y comprendan los datos. Desde la perspectiva institucional, permite identificar datasets con potencial para nuevas investigaciones, colaboración interdisciplinaria, Inteligencia Artificial e integración con otros conjuntos de datos.

A futuro, la acumulación ordenada de estos registros permitirá avanzar hacia un Catálogo Institucional de Conjuntos de Datos de Investigación de la UDES. Cada proyecto aportará un dataset y su documentación; colectivamente, los proyectos contribuirán a construir un patrimonio institucional de datos de investigación.

Por lo anteriormente expuesto, esta guía orienta sobre la documentación del conjunto de datos, su contenido, formato y medio de organización en GitHub; plataforma en la nube para el almacenamiento.

4. Estructura general del repositorio GitHub

Cada proyecto aprobado tendrá una carpeta propia en la raíz del repositorio, identificada como PROYECTO_<nombre-corto> (ejemplo: PROYECTO_PredAgro). Dentro de ella se utilizará la siguiente estructura estándar.

Carpeta	Contenido
README.md	Resumen ejecutivo y ficha del proyecto (Anexo 4).
01-propuesta/	Propuesta aprobada en la convocatoria.
02-documento-tecnico/	Documento Técnico del Conjunto de Datos (Anexo 1).
03-diccionario-datos/	Diccionario de variables o dimensiones del dataset y metadatos
04-dataset-publicado/	Información y enlace al dataset publicado; no contiene el dataset.
05-cumplimiento-fair/	Checklist de cumplimiento de los principios FAIR (Anexo 2).



06-consideraciones-éticas/	Soportes éticos y legales pertinentes.
07-productos-adicionales/	Informes, artículos, ponencias y otros productos asociados.

5. Documentación del conjunto de datos

Para esta convocatoria, la documentación del conjunto de datos debe permitir que una persona externa al equipo investigador pueda identificar, comprender e interpretar el dataset. Por ello, no se limita al archivo de datos.

Como mínimo, la documentación debe integrar la descripción general del dataset, sus metadatos, el diccionario de datos y muestras ejemplificadas y documentadas.

Ejemplo de metadatos: título del dataset, autores, institución, proyecto de origen, descripción, palabras clave, área de conocimiento, cobertura temporal y geográfica, fecha, versión, formato, licencia, condiciones de acceso, repositorio y DOI entre otros. Esta información corresponde a los metadatos incluidos en el repositorio del dataset al momento de su publicación y de acuerdo a los requisitos del mismo.

El dataset de cada proyecto NO se alojará en el repositorio institucional de GitHub. El equipo investigador deberá publicarlo en un repositorio de datos reconocido, de acuerdo con las condiciones aplicables al conjunto de datos. En GitHub se conservará la información necesaria para localizarlo y comprender sus condiciones de acceso.

En la carpeta 04-dataset-publicado/ se incorporará únicamente el archivo ENLACE_DATASET.md, con la siguiente información: nombre del dataset, repositorio y URL de acceso, identificador persistente (DOI), licencia, fecha de publicación, formato y condiciones de acceso.

Ejemplo de información del diccionario de datos: nombre de la variable, descripción, tipo de dato, unidad de medida, valores permitidos, codificación, valores faltantes y observaciones.



6. Formatos aceptados para los archivos en cada carpeta

Carpeta	Formatos	Orientación
01-propuesta/	.pdf, .docx	Versión final aprobada.
02-documento-tecnico/	.pdf, .docx, .md	Debe cubrir las secciones definidas en el Anexo 1.
03-diccionario-datos/	.csv, .xlsx	Archivo independiente cuando su extensión lo requiera.
04-dataset-publicado/	.md	Solo el archivo de enlace; nunca datos crudos.
05-cumplimiento-fair/	.md, .pdf, .docx	Checklist FAIR.
06-consideraciones-eticas/	.pdf	Soportes pertinentes.
07-productos-adicionales/	.pdf, .docx, .pptx	Productos asociados.

7. Convención de nombres de archivo

Se recomienda utilizar la estructura: <CODIGO-PROYECTO>_<tipo-documento>_v<AAAAMMDD>.<ext>

Ejemplo: PredAgro_documento-tecnico_v20260805.docx

8. Checklist de cumplimiento FAIR (Anexo 2)

Dimensión	Preguntas orientadoras
F – Encontrables	¿Tiene DOI o identificador persistente? ¿Está publicado en un repositorio reconocido? ¿Tiene título, descripción y metadatos claros?
A – Accesibles	¿Se puede acceder al dataset o solicitar acceso? ¿Están claras las condiciones de uso? ¿Existe una licencia? ¿La documentación es accesible?
I – Interoperables	¿Utiliza formatos abiertos y estándares? ¿Las variables emplean vocabularios o clasificaciones pertinentes? ¿Existe diccionario de datos?
R – Reutilizables	¿La documentación permite comprender y reutilizar el dataset? ¿Se describen origen, curación y limitaciones? ¿Están claras las condiciones éticas, legales y de licencia?



9. Proceso de entrega

Crear la carpeta PROYECTO_<nombre-corto> en la raíz del repositorio, a partir de la plantilla estándar.

2. Diligenciar el README.md de la carpeta con la ficha resumen del proyecto.

3. Cargar cada entregable en la subcarpeta correspondiente, respetando los formatos aceptados.

4. En 04-dataset-publicado/, completar únicamente ENLACE_DATASET.md — nunca subir el dataset.

5. Diligenciar el checklist FAIR de la carpeta 05-cumplimiento-fair/.

6. Comprimir carpeta y adjuntar como una evidencia en el SIGIIP.

10. Del repositorio al Catálogo Institucional

La estructura común permitirá que los productos de los diferentes proyectos puedan organizarse posteriormente como un catálogo institucional. El catálogo podrá registrar, entre otros, el nombre e identificador del dataset, proyecto de origen, grupo e investigador, área de conocimiento, palabras clave, tipo de datos, cobertura, año, estado de publicación, nivel de acceso y enlaces a su documentación.

La finalidad es que la Universidad pueda identificar qué conjuntos de datos posee, cuáles están disponibles, cuáles tienen potencial de reutilización, cuáles pueden integrarse con otros datos y cuáles pueden generar nuevas investigaciones o desarrollos tecnológicos.

La propuesta debe entenderse como una estrategia progresiva: cada proyecto organiza correctamente sus productos y, con la acumulación de proyectos, la institución construye un catálogo común.

11. Orientación para la toma de decisiones institucionales

La documentación organizada también permitirá que la Vicerrectoría de Investigaciones valore posteriormente el potencial de los datasets y tome decisiones sobre publicación, actualización, integración, priorización, nuevas investigaciones y desarrollos basados en datos.



Información disponible	Posible decisión institucional
Calidad y nivel de documentación	Definir necesidad de fortalecimiento o curaduría adicional.
Potencial de IA	Identificar datasets para futuras investigaciones o desarrollos.
Potencial interdisciplinario	Promover colaboración entre grupos y áreas.
Posibilidad de integración	Identificar oportunidades de combinar datasets.
Nivel de acceso y licencia	Definir estrategias de divulgación y reutilización.
Potencial de continuidad	Considerar actualización o nuevas fases de investigación.

12. Orientación final para los equipos de investigación

El conjunto de datos debe entenderse como un producto de investigación y su documentación como parte de su valor científico. Entregar un dataset no significa únicamente entregar un archivo; significa proporcionar la información necesaria para que otros investigadores puedan comprender qué contiene, de dónde proviene, cómo fue construido, cómo está estructurado, bajo qué condiciones puede utilizarse y qué posibilidades futuras ofrece.

La organización propuesta permitirá que los productos de los proyectos de la convocatoria sean progresivamente integrados en un Catálogo Institucional de Conjuntos de Datos de Investigación de la UDES, fortaleciendo la ciencia abierta, la visibilidad de la investigación, la colaboración interdisciplinaria y el aprovechamiento futuro de los datos.