

Datos FAIR en Argentina

Desarrollo y desafíos del Repositorio de Datos Académicos RDA-UNR



Paola Carolina Bongiovani

Universidad Nacional de Rosario. Facultad de Humanidades y Artes. Centro de Estudios Interdisciplinarios en Bibliotecología y Ciencias de la Información, Argentina / pbongio@unr.edu.ar | <https://orcid.org/0000-0003-0049-9086>

Fernando Díaz Pacífico

Universidad Nacional de Rosario. Facultad de Humanidades y Artes. Centro de Estudios Interdisciplinarios en Bibliotecología y Ciencias de la Información, Argentina / fdiazpacifico@fbiof.unr.edu.ar | <https://orcid.org/0000-0002-6491-590X>

Paulina Freán

Universidad Nacional de Rosario. Facultad de Humanidades y Artes. Centro de Estudios Interdisciplinarios en Bibliotecología y Ciencias de la Información, Argentina / paulina.frean@unr.edu.ar | <https://orcid.org/0000-0002-9710-0213>

Resumen

El propósito de este artículo es comunicar los resultados de la implementación de un repositorio de datos abiertos de investigación para la Universidad Nacional de Rosario (UNR). En primer término, se explica el proceso de evaluación, selección, y puesta en funcionamiento del Repositorio de Datos Académicos UNR (RDA-UNR), utilizando el software Dataverse, desarrollado por la Universidad de Harvard. A continuación, se describen las principales actividades desarrolladas por los profesionales de la información en el marco de un trabajo interdisciplinario con investigadores de la UNR. El RDA-UNR fue planificado para organizar, preservar, diseminar y dar visibilidad a los datos de investigación producidos por los investigadores de la UNR, adoptando principios FAIR y estándares internacionales en el marco de la Ciencia Abierta. Resultados: El RDA-UNR fue inaugurado en agosto de 2022, sirve a una comunidad de investigación de 12 unidades académicas, 13 institutos de doble dependencia UNR-CONICET; cuenta con 66 datasets curados y 38 en proceso de curación, siendo el primer repositorio en su tipo de Argentina.

Palabras clave

Ciencia abierta
Repositorios de datos
Datos abiertos
Principios FAIR

FAIR Data in Argentina: Development and challenges of the Academic Data Repository RDA-UNR

Abstract

The purpose of this article is to present the results of the implementation of an open research data repository for the Universidad Nacional de Rosario (UNR). First, the article explains the process of evaluating, selecting, and deploying the Academic Data Repository UNR (RDA-UNR) using the Dataverse software developed by Harvard University. Secondly, the main activities undertaken by information professionals

Keywords

Open science
Data repositories
Open data
FAIR Principles

in collaboration with UNR researchers within an interdisciplinary framework are described. The RDA-UNR was specifically designed to organize, preserve, disseminate, and enhance the visibility of research data generated by UNR researchers, adhering to FAIR principles and international standards aligned with the Open Science framework. Results: The RDA-UNR was officially launched in August 2022 and now supports a research community comprising 12 academic units and 13 UNR-CONICET joint institutes. It includes 66 curated datasets as well as 38 in the process of being curated, making it the first repository of its kind in Argentina.

Artículo recibido: 18-12-2024. Aceptado: 30-05-2025.

Ciencia abierta, datos de investigación y repositorios¹

1. Este artículo es una versión revisada, adaptada y abreviada del capítulo 5 de la tesis doctoral (Bongiovani, 2024) titulada "Factores individuales e institucionales que influyen en la intención de compartir datos de investigación en repositorios. El caso de la Universidad Nacional de Rosario", presentada en la Universidad Carlos III de Madrid.

La investigación y la innovación científica son el pilar fundamental en la búsqueda de soluciones para los complejos desafíos globales. Para enfrentarlos, hoy en día disponemos de una cantidad sin precedentes de datos y avances tecnológicos que no solo permiten el almacenamiento y procesamiento de datos, sino que también proporcionan herramientas cada vez más poderosas, como la Inteligencia Artificial, para analizar esta información. Además, la Web y las herramientas relacionadas facilitan la comunicación instantánea entre científicos de todo el mundo, promoviendo la colaboración y la distribución inmediata y abierta del conocimiento generado. Se espera que la ciencia basada en datos acelere la investigación y profundice sus impactos, aportando beneficios tangibles para la sociedad en su conjunto.

Este nuevo paradigma está cambiando el modo de hacer ciencia (Anglada y Abadal, 2018) e implica mayor transparencia, rigor, responsabilidad y la posibilidad de reproducir las investigaciones, mostrando todo el proceso de investigación, no solamente los resultados finales sino también cómo se llegó a esos resultados. El eje principal es el impacto social que se quiere lograr con la ciencia abierta (Méndez Rodríguez, 2021).

La ciencia abierta es definida como "conocimiento transparente y accesible que se comparte y desarrolla a través de redes colaborativas" (Vicente-Saez y Martínez-Fuentes, 2018). Entre los componentes de la ciencia abierta, se encuentran los datos de investigación abiertos, sobre los que se centrará este trabajo. Específicamente el caso del Repositorio de Datos Académicos de la Universidad Nacional de Rosario (RDA-UNR).

No existe una definición universalmente aceptada de "datos de investigación". Instituciones internacionales como el National Institute of Health (NIH), definen los datos científicos como aquellos materiales registrados y aceptados comúnmente por la comunidad científica, cuya calidad es suficiente para validar y replicar los resultados de la investigación, independientemente de si estos datos se utilizan para respaldar publicaciones académicas. Esta definición también especifica lo que no se incluye en ella (NIH, 2020).

En el marco del RDA-UNR trabajamos con una definición operacional amplia, como se expresa en el sitio web RDA-UNR Info: "Los datos de investigación son información recolectada, observada o generada durante el proceso de investigación, que fundamenta las afirmaciones de la investigación y sirve para validar los resultados" (Universidad Nacional de Rosario, s.f.). La legislación argentina también provee una definición en esta línea (Resolución del Ministerio de Ciencia, Tecnología e Innovación Productiva 753 - E/2016).

No cabe duda de que los datos son una fuente primaria esencial para respaldar la investigación científica y permitir la realización de descubrimientos tanto básicos como aplicados o teóricos, constituyendo productos fundamentales en los proyectos de investigación.

Para que los estudios sean replicables y reproducibles, los investigadores y curadores de datos siguen las recomendaciones y buenas prácticas cristalizadas en cuatro principios, denominados Principios FAIR (Findable, Accessible, Interoperable, Reusable) (Wilkinson et al., 2016; Force11, 2020), con el objetivo de hacer los datos tan accesibles como sea posible. No obstante, reconocen que en algunos casos es necesario restringir el acceso a los datos debido a limitaciones relacionadas con la seguridad, privacidad, confidencialidad, entre otras restricciones.

La participación de los profesionales de la información en este campo ha sido ampliamente documentada y valorada (Brochu y Burns, 2019; Pinfield, Cox y Smith, 2014), influyendo tanto en la creación de políticas e infraestructuras, como en la prestación de servicios especializados.

La Universidad Nacional de Rosario

La Universidad Nacional de Rosario es una de las principales universidades nacionales públicas de Argentina, en cuanto a su producción científica. Fue una de las primeras en desarrollar su repositorio institucional RepHipUNR “para archivar, organizar, preservar y distribuir digitalmente, en variados formatos, tanto materiales de enseñanza y aprendizaje como la producción científica de Investigación y Desarrollo (I+D) de los profesores, profesionales e investigadores de la UNR” (Bongiovani et al., 2014a; San Martín et al., 2010).

A partir de 2016 este trabajo se fue formalizando en políticas institucionales: Resolución de Consejo Superior N° 1842/2016 (UNR, 2016), Política Institucional de Acceso Abierto de la Universidad Nacional de Rosario (Resolución 5425/2017), Unidad de Gestión de Acceso Abierto (Resolución 2159/2018. Unidad de Gestión de Acceso Abierto).

En la Política Institucional de Acceso Abierto (PIAA) se establece que, de acuerdo con la Ley 26.899, los datos primarios generados en procesos de investigación, financiados con fondos del Estado deben estar a disposición abiertamente en el repositorio institucional en un plazo no mayor a cinco años a partir del momento de su recolección (Bongiovani, Voras y Pérez Bigot, 2019b).

En 2020, la UNR adoptó la Agenda 2030 como herramienta de planificación para su desarrollo, la cual incluye, entre otros aspectos, la construcción de un sistema científico de valor público que promueve el acceso abierto a los resultados de la investigación científica.

Antes de la creación de la PIAA, y para entender la situación en la UNR en relación con la gestión de datos primarios de investigación, se realizó un análisis de las necesidades (Bongiovani y Martínez-Uribe, 2014b) y se encontró que los investigadores mayormente compartían sus datos de manera controlada, entre colegas conocidos a través de correo electrónico, descarga en sitios web y compartiendo dispositivos, no de manera abierta.

Este mismo estudio de 2014 detectó la necesidad de generar infraestructuras y servicios para satisfacer necesidades de gestión de datos de investigación. Los investigadores que contestaron una pregunta específica sobre servicios (n=375) indicaron que estaban interesados en contar con asesoramiento sobre cuestiones prácticas relacionadas con la gestión de datos (64%), sobre opciones para almacenar, gestionar y compartir datos de forma segura (63%), sobre preservación de datos de investigación (55%) y sobre digitalización, herramientas y servicios (53%) entre otros.

Se puso en evidencia la necesidad de implementar un repositorio específico para compartir datos de investigación que cumpla con estándares internacionales. Los datos de investigación difieren en muchos aspectos respecto de las publicaciones (tesis, artículos, libros, material didáctico), y requieren una gestión específica, dada la variedad de datos y la granularidad necesaria en la descripción (metadatos adaptados a las áreas temáticas de los datos) entre otros elementos (Hernández Pérez y García-Moreno, 2013) que hacen al cumplimiento de los principios de datos FAIR para que los datos se puedan encontrar, sean accesibles, interoperables y reutilizables.

En 2021, durante el proceso de planificación e implementación del RDA-UNR, se envió una encuesta a través de las Secretarías de Ciencia y Tecnología de la Universidad, a fin de identificar necesidades en relación con los datos de investigación para poder planificar la infraestructura y políticas adecuadas para su gestión. Aunque la tasa de respuesta fue muy baja, solamente 3,35% respondieron de los 3521 investigadores de la UNR (RACT, 2021) allí se encontró que el 86% de los investigadores que respondieron la encuesta (n=118) no habían depositado datos en repositorios de datos o servicios similares.

La elección de Dataverse

Se realizó una evaluación de los paquetes de software existentes para repositorios de datos, su adopción a nivel global y en Latinoamérica. Se decidió utilizar el software libre del Proyecto Dataverse que actualmente cuenta con 125 instalaciones alrededor del mundo incluyendo instituciones de Brasil, Chile, Perú, Colombia, Ecuador y México (Dataverse Project, s.f.).

El software Dataverse fue desarrollado originalmente por el Instituto de Ciencias Sociales Cuantitativas (IQSS) de la Universidad de Harvard en 2006. Desde entonces, se ha consolidado como una herramienta clave para la gestión, publicación y preservación de datos de investigación. Su diseño modular y jerárquico permite adaptarse a distintas necesidades institucionales, disciplinarias y de investigación, promoviendo el cumplimiento de los principios FAIR (Encontrable, Accesible, Interoperable, Reutilizable).

Dataverse organiza la información en colecciones jerárquicas, denominadas también “dataverses”. Cada colección puede contener datasets (conjuntos de datos) y, a su vez, otras subcolecciones, lo que permite estructurar los datos por unidades temáticas, proyectos, grupos de investigación, departamentos o instituciones completas. Esta organización flexible facilita la asignación diferenciada de permisos de acceso, esquemas de metadatos y términos de uso en cada nivel.

El funcionamiento básico de Dataverse puede entenderse en los siguientes pasos:

1. Creación de una colección o subcolección (Dataverse): Los usuarios con permisos adecuados pueden crear colecciones en las que se definan políticas de acceso, tipos de metadatos requeridos, términos de uso y parámetros de publicación.
2. Carga y descripción de datasets: Los investigadores pueden subir conjuntos de datos a estas colecciones. Cada dataset puede incluir diversos tipos de archivos (por ejemplo, bases de datos, documentos, imágenes, scripts), a los que se les asocian metadatos que describen su contenido, autoría, licencia y condiciones de uso.

3. Asignación automática de citas e identificadores persistentes: Cada conjunto de datos publicado genera automáticamente una cita académica con un identificador persistente (DOI en el caso de UNR, pero se puede configurar también Handle). Incluso cada archivo dentro del dataset puede tener su propio identificador. Esto garantiza la trazabilidad y la citabilidad de los datos.
4. Gestión de metadatos: Dataverse permite trabajar con distintos esquemas de metadatos (como Dublin Core, DDI, DataCite, OpenAIRE o Schema.org), que pueden configurarse según las necesidades de cada colección. Esta flexibilidad facilita la interoperabilidad entre sistemas y la adaptación a diferentes disciplinas científicas.
5. Opciones de control de acceso: Aunque Dataverse está diseñado para facilitar el acceso abierto a los datos, permite aplicar restricciones cuando sea necesario. Existen tres niveles de acceso:
 - » Acceso público con aceptación de términos de uso: Los archivos son accesibles tras aceptar condiciones específicas definidas por el autor o institución.
 - » Archivos restringidos: Parte del dataset está disponible, pero ciertos archivos solo pueden ser descargados por usuarios autorizados.
 - » Dataset completamente restringido: Aunque los metadatos permanecen públicos e indexados, ni los datos ni su descripción completa están disponibles para descarga.
6. Preservación y compatibilidad de formatos: En el caso de archivos tabulares (como los utilizados en SPSS, STATA o CSV), Dataverse realiza un procesamiento especial: separa metadatos variables (nombres, etiquetas, tipos de datos), genera estadísticas resumen por variable, y guarda los archivos en un formato de preservación independiente del software original. Además, conserva el archivo original y genera una huella digital (UNF - Universal Numerical Fingerprint) que permite verificar que los datos no fueron alterados.
7. Visualización y búsqueda avanzada: Los usuarios pueden buscar y navegar datasets dentro de una colección o en todo el repositorio, utilizando filtros sobre los campos de metadatos. También es posible visualizar datos tabulares directamente desde la interfaz.
8. Interoperabilidad técnica: Dataverse implementa el protocolo OAI-PMH, lo que permite su integración con otros sistemas y buscadores académicos. Además, ofrece una variedad de APIs (Application Programming Interfaces) para facilitar la interacción con otras plataformas y servicios digitales.
9. Configuración de políticas y términos de uso: Cada colección o dataset puede establecer términos de uso personalizados que los usuarios deben aceptar para acceder o publicar datos, mediante un mecanismo de clic (“click-through license”).

Gracias a este conjunto de funcionalidades, Dataverse facilita el cumplimiento de buenas prácticas para el acceso abierto, la preservación y la reutilización de datos, en la medida en que se combine con políticas institucionales claras y una infraestructura técnica adecuada. Su diseño modular y configurable permite adaptarse a distintos entornos institucionales y disciplinares. Las capacidades del software se describen en mayor detalle en su sitio oficial y en la documentación técnica disponible (Crosas, 2011; Crosas, 2019).

Desde 2021, la Universidad se unió como miembro del Global Dataverse Community Consortium (GDCC) y, a través de GDCC, es miembro también de DataCite para gestionar los Digital Object Identifiers (DOI) del repositorio.

El repositorio se encuentra registrado en el directorio re3data.org con su DOI correspondiente: <http://doi.org/10.17616/R31NJN9K>.

Asimismo, fue registrado en FAIRsharing.org con el DOI: <https://doi.org/10.25504/FAIRsharing.d81f99>.

El proceso de construcción del Repositorio de Datos Académicos RDA-UNR

El Repositorio de Datos Académicos de la Universidad Nacional de Rosario (RDA-UNR) permite publicar y acceder a los datos de investigación generados por la comunidad de la UNR. Además, los investigadores de otras instituciones que publican artículos en revistas del Portal de Revistas UNR pueden depositar los datos que sustentan las investigaciones de sus artículos publicados.

El objetivo del repositorio es dar visibilidad, potenciar la reutilización y asegurar la preservación de los datos, dando cumplimiento al estándar de datos FAIR.

La Universidad, a través del Comité de Acceso Abierto y Unidad de Gestión de Acceso Abierto contrató el asesoramiento experto de la Dra. María Esteva (Texas Advanced Computing Center, The University of Texas, Austin) desde junio de 2021. En la primera reunión participaron la coordinadora de la Unidad de Gestión de Acceso Abierto (UGAA) y su equipo de trabajo (dos bibliotecarias y una administradora de sistemas), dos bibliotecarios y dos investigadores del Área Sensores Remotos (Escuela de Agrimensura de la Facultad de Cs. Exactas, Ingeniería y Agrimensura). Esta área había mostrado alto interés en la capacitación anterior y mediante un intercambio fluido de comunicaciones con la coordinación de la UGAA.

Se propuso un programa de trabajo para la planificación, implementación y evaluación del repositorio. Se acordó un plan que contemplara: la selección de la tecnología de implementación del repositorio (Dataverse) y demás aspectos necesarios, infraestructura institucional, políticas y buenas prácticas, servicios, capacitación a investigadores, estrategias de comunicación y evaluación. Para ello fue necesario formar un equipo de trabajo multidisciplinario y obtener casos de datos para aprender y enseñar la curación y publicación de datos. Hasta la fecha de lanzamiento del RDA UNR, se llevaron a cabo más de 50 reuniones de trabajo interdisciplinarias y de formación².

En esta etapa, bibliotecarios curadores trabajaron junto a los investigadores para conocer las características del proceso de investigación, lo que permitió crear esquemas para identificar datos y tipos de archivos (Fig. 1).

En las siguientes reuniones, con una agenda concreta de trabajo para cada grupo, todos los participantes se comprometieron a testear el software Dataverse instalado. Se comenzó a trabajar con datos de proyectos de investigación concretos. En algunos casos, se trabajó con datos de artículos ya publicados, porque les daba a los investigadores la tranquilidad para compartir datos que ellos ya habían explotado.

2. Se agradece a la Dra. Larisa Cybulski y Dra. Ana Bortolotti (FBIOyF); Mg. Laura Rita Balparda, TCS Diego López (FCEIA); Lic. Estefanía Aranda. (FCPOLIT); Dra. Natalia Santucci y Dra. Mariana Lagrutta (FCM) por su participación en los primeros procesos de curación de datos.

Flujo de trabajo: 1ª ETAPA



Figura. 1. Ejercicio de identificación de datos producidos en un estudio de diseño de proteínas y biología molecular en el proceso de curación de los datos. Fuente: Elaboración por Larisa Cybulski, Ana Bortolotti y Fernando Díaz Pacífico.

En esta etapa también se definió el nombre del repositorio. Se empezó a explorar una herramienta wiki (MediaWiki) para compartir la información del equipo de trabajo y las políticas, buenas prácticas, entre otros documentos. Sin embargo, la herramienta fue considerada compleja de manejar por el equipo por tener que aprender a marcar el texto. Se decidió instalar un sitio WordPress que acompaña al repositorio con toda la información relacionada. Se estableció una cuenta de correo institucional repositoriodatos@unr.edu.ar y una cuenta en Google (repositoriodatosunr@gmail.com) donde se comparten los materiales de trabajo interno del equipo (minutas reuniones, presentaciones en eventos, capacitaciones, marketing, etc.).

RDA-UNR Organización y Estructura

El RDA-UNR funciona bajo la órbita de la Unidad de Gestión de Acceso Abierto, Comité de Acceso Abierto de la UNR. En octubre de 2023, la UNR definió que esta unidad depende de la Coordinación de Gabinete, un área que “promueve y desarrolla políticas integrales y articuladas de la Universidad”.

El comité de Acceso Abierto está compuesto por las siguientes áreas de gestión de la Universidad: Coordinación de Gabinete de Rectorado, Área Académica y de Aprendizaje, Área de Ciencia, Tecnología e Innovación para el Desarrollo, Consejo de Investigaciones CIUNR y Centro de Estudios Interdisciplinarios.

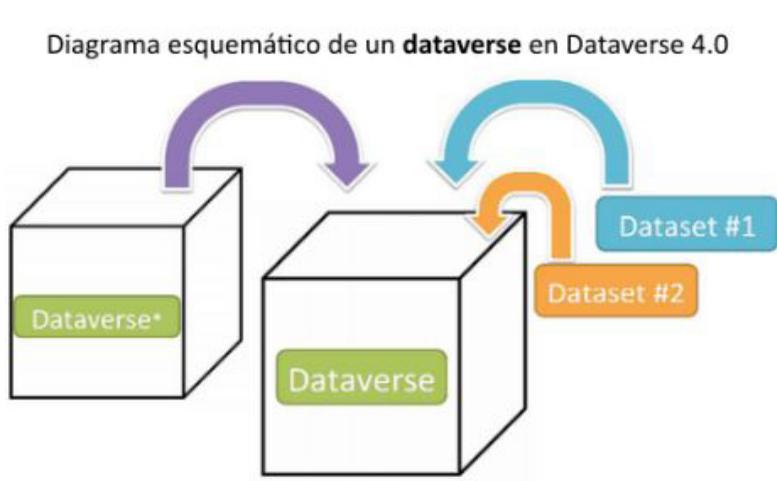
Si bien existe una normativa sobre la forma de coordinación entre la Unidad de Gestión de Acceso Abierto y las unidades académicas (Resolución 2159/2018) por la cual las facultades fueron designando equipos de trabajo para el repositorio institucional de publicaciones, en el caso del RDA-UNR la coordinación es dispar y muchas no han designado aún su equipo de trabajo específico para el repositorio de datos. Dada la variedad de disciplinas en las diferentes facultades, se está trabajando en la conformación de un Consejo Asesor para el repositorio.

El repositorio está compuesto de una instalación de Dataverse <https://dataverse.unr.edu.ar/> donde se suben los conjuntos de datos y un sitio de información RDA-UNR Info <https://dataverse-info.unr.edu.ar/> desarrollado en WordPress.

El repositorio está organizado en diferentes colecciones (llamadas Dataverses) correspondientes a facultades y al portal Revistas, además de la colección para Unidades Ejecutoras de Doble Dependencia UNR-CONICET.

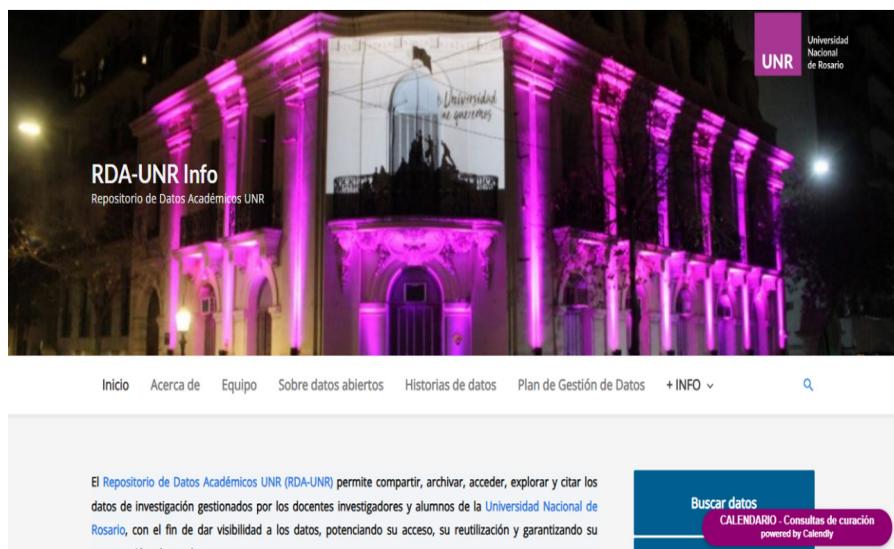
Como se observa en el diagrama de la figura 2, el contenido del repositorio —compuesto por archivos de datos y metadatos— se organiza en colecciones denominadas dataverses, que en el caso del RDA-UNR se corresponden, en términos generales, con las unidades académicas o facultades de la universidad. Cada dataverse alberga distintos conjuntos de datos (datasets), y el sistema permite, además, la creación de subcolecciones dentro de cada uno. Esta funcionalidad responde a necesidades específicas de la comunidad académica, como brindar visibilidad a la producción científica de un centro de investigación, una cátedra o un laboratorio.

Figura 2. Diagrama de colecciones y estructura jerárquica en Dataverse. Guía de Uso de Dataverse.



Fuente: Guía de Uso de Dataverse <https://dataverse-info.unr.edu.ar/guia-de-uso/>

Figura 3. Página de inicio del Dataverse en RDA-UNR.



Fuente: <https://dataverse.unr.edu.ar/>

Dentro de cada Dataverse se pueden crear otros Dataverses para agrupar 5 o más conjuntos de datos (datasets) correspondientes a proyectos, laboratorios, centros de investigación y otras unidades dentro de cada Facultad.

La Figura 4 representa esquemáticamente la estructura de un dataset en Dataverse. Un dataset funciona como un contenedor que agrupa tres tipos principales de archivos: archivos de datos (por ejemplo, bases de datos en formato tabular), documentación (como cuestionarios, protocolos, diccionario de datos, libro de códigos, archivos README, entre otros) y código (scripts utilizados en el análisis o procesamiento de datos). Además de estos componentes, cada dataset está acompañado por un conjunto de metadatos descriptivos, que permiten identificar, contextualizar, enlazar y recuperar la información mediante buscadores y sistemas interoperables. Esta estructura modular permite garantizar una presentación coherente y completa del material de investigación, promoviendo la reutilización y facilitando la comprensión de los datos por parte de otros usuarios.

Actualmente el repositorio cuenta con 66 datasets publicados y 38 datasets en proceso de curación.

Diagrama esquemático de un **dataset** en Dataverse 4.0

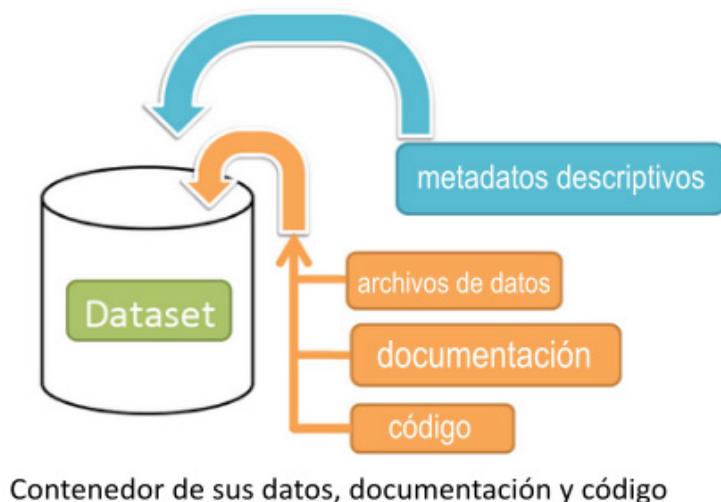


Figura 4. Esquema de un dataset como un contenedor de datos, documentación, código y metadatos en Dataverse.

Fuente: Guía de Uso de Dataverse <https://dataverse-info.unr.edu.ar/guia-de-uso/>

El equipo de trabajo es interdisciplinario y está compuesto por profesionales de la información, administradores de sistemas e investigadores (Bongiovani et al., 2022). Algunos bibliotecarios han participado activamente como curadores en el proceso de implementación del repositorio y, en la etapa posterior, como curadores de área temática.

Políticas y directrices

El repositorio se rige por políticas establecidas durante su planificación. Para el desarrollo e implementación, se tuvieron en cuenta las necesidades de su comunidad de investigadores que se evidenciaron principalmente en las consultas de curación

y reuniones del equipo. Allí se detallan los criterios para el uso, publicación y preservación de los datos, así como los roles y responsabilidades de quienes participan en el proceso: autores, curadores, administradores y usuarios.

A continuación, se presentan los aspectos más relevantes para comprender el marco de funcionamiento del repositorio:

- » Finalidad y alcance: El RDA-UNR está destinado a la gestión y publicación de datos de investigación generados en el ámbito de la Universidad Nacional de Rosario, o en trabajos de autores que publican en revistas de la Universidad.
- » Responsabilidad de los autores: Al publicar datos, los usuarios deben garantizar la autoría y licenciamiento adecuado, evitar la inclusión de información sensible o confidencial, y aceptar que una vez asignado un DOI, las modificaciones implicarán una nueva versión del dataset manteniendo su DOI.
- » Licencia predeterminada: El repositorio adopta por defecto la licencia Creative Commons Atribución 4.0 Internacional (CC BY 4.0), aunque es posible seleccionar otras opciones según el caso.
- » Citación de datos: Se promueve activamente la citación de los datasets reutilizados en trabajos académicos, incluyendo su identificador persistente (DOI).
- » Privacidad y tratamiento de datos personales: Se resguarda la información de los usuarios y se cumple con la Ley 25.326.
- » Datos aceptados: Se recomienda utilizar formatos abiertos, sin restricciones de tipo o tamaño, aunque los datasets más extensos requieren un trabajo de curación detallado.
- » Preservación: La infraestructura y la política de preservación están a cargo del Área de Infraestructura Tecnológica de la UNR.

Para más detalles sobre políticas específicas, instrucciones para autores, lineamientos técnicos o aspectos de curación, puede consultarse la sección correspondiente del sitio de información del RDA-UNR <https://dataverse-info.unr.edu.ar/>

Los principios FAIR en los conjuntos de datos del RDA-UNR

El cumplimiento de los principios FAIR en el software Dataverse ha sido documentado (Crosas, 2020, Conzett, 2020). Se desarrolla su implementación en el RDA-UNR:

F. Para ser Localizables (Findable)

F1. A los (meta)datos[1] se les debe asignar un identificador único y persistente.

El software Dataverse implementa F1 con el uso de identificadores persistentes (PID) como DOI, Handle y ORCID. La asignación de un PID a nivel de conjunto de datos es obligatoria y asignar un identificador persistente a nivel de cada archivo es opcional. En el caso del repositorio RDA-UNR se usa la versión 5.10.1 del software Dataverse. Desde el inicio del repositorio se contrató el servicio DataCite a través del consorcio Dataverse. Se asignan DOIs a todos los conjuntos de datos (figura 5), así como a cada uno de los archivos (figura 6). Además, la mayoría de los autores agregan su identificador ORCID (figura 7) al conjunto de datos (Bongiovani et al., 2023; Salazar et al., 2023).

RDA-UNR | Repositorio de Datos Académicos

Buscar ▾ Guía de uso Soporte Español ▾ Registrarse Iniciar sesión

RDA UNR > Revistas UNR > Cuadernos del CURIHAM >

Datos de: Modelación matemática hidrológica-hidráulica del escurrimiento superficial en la cuenca del A° Pavón (Santa Fe, Argentina)

Versión 1.0

Stenta, Hernán; Gerardo Riccardi; Pedro Basile; Carlos Scuderi, 2022. "Datos de: Modelación matemática hidrológica-hidráulica del escurrimiento superficial en la cuenca del A° Pavón (Santa Fe, Argentina)", <https://doi.org/10.57715/UNR/THCOKS>, RDA UNR, V1, UNF:6:iSPsB46f4Myp3OVdezelg== [fileUNF]

Acceder al dataset ▾
 Contactar al propietario Compartir

Métricas del dataset
 879 Views
 100 Downloads
 1 Citation

Descripción

El objetivo principal de la investigación es avanzar en la descripción del comportamiento del escurrimiento superficial y su impacto en cuencas del sur de la provincia de Santa Fe, en este caso se realiza el estudio sobre la cuenca del A° Pavón (sur de la provincia de Santa Fe).

El objetivo de los datos recolectados fue utilizarlos para la constitución, calibración y explotación de un modelo hidrológico-hidráulico que permite obtener alturas máximas de

Figura 5. Asignación de identificadores persistentes DOI a nivel de dataset (RDA-UNR).

Fuente: Captura de pantalla de <https://dataverse.unr.edu.ar/> Stenta et al., 2022

RDA-UNR | Repositorio de Datos Académicos

Buscar ▾ Guía de uso Soporte Español ▾ Registrarse Iniciar sesión

RDA UNR > Revistas UNR > Cuadernos del CURIHAM >

Datos de: Modelación matemática hidrológica-hidráulica del escurrimiento superficial en la cuenca del A° Pavón (Santa Fe, Argentina) >

HMAX_ENERO_2017.tif

Este archivo es parte de "Datos de: Modelación matemática hidrológica-hidráulica del escurrimiento superficial en la cuenca del A° Pavón (Santa Fe, Argentina)".

Versión 1.0

Citación de archivo

Stenta, Hernán; Gerardo Riccardi; Pedro Basile; Carlos Scuderi, 2022. "HMAX_ENERO_2017.tif", Datos de: Modelación matemática hidrológica-hidráulica del escurrimiento superficial en la cuenca del A° Pavón (Santa Fe, Argentina), <https://doi.org/10.57715/UNR/THCOKS/MOIONR>, RDA UNR, V1

Acceder al archivo ▾
 Contactar al propietario Compartir

Métricas del archivo
 10 Downloads

Citación del dataset

Stenta, Hernán; Gerardo Riccardi; Pedro Basile; Carlos Scuderi, 2022. "Datos de: Modelación matemática hidrológica-hidráulica del escurrimiento superficial en la cuenca del A° Pavón (Santa Fe, Argentina)", <https://doi.org/10.57715/UNR/THCOKS>, RDA UNR, V1, UNF:6:iSPsB46f4Myp3OVdezelg== [fileUNF]

Citar dataset ▾ Obtenga información sobre Estándares de cita de datos.

Figura 6. Asignación de DOI al archivo dentro del conjunto de datos (RDA-UNR).

Fuente: Captura de pantalla de <https://dataverse.unr.edu.ar/> Stenta et al., 2022

RDA-UNR | Repositorio de Datos Académicos

Buscar ▾ Guía de uso Soporte Español ▾ Registrarse Iniciar sesión

Identificador persistente del dataset doi:10.57715/UNR/EXIVRO

Fecha de publicación 2023-08-01

Título Datos de: Huella Urbana de la Ciudad de Rosario, Santa Fe, Argentina

Título alternativo Dataset: Urban Footprint of the City of Rosario, Santa Fe, Argentina

Autor

Balparda, Laura Rita (Área de Sensores Remotos. Escuela de Agrimensura. FCEIA. UNR) - ORCID: 0000-0002-7666-8213

del Valle, Héctor (Centro Nacional Patagónico, Consejo Nacional de Investigaciones Científicas y Técnicas) - ORCID: 0000-0002-3329-1057

López, Diego Alejandro Germán (Área de Sensores Remotos. Escuela de Agrimensura. FCEIA. UNR) - ORCID: 0000-0003-0639-4613

Torralba, María Cecilia (Escuela de Agrimensura. FCEIA. UNR) - ORCID: 0000-0003-0953-2444

Tazzioli, Florencia (Estudiante, Escuela de Agrimensura. FCEIA. UNR) - ORCID: 0000-0002-5379-7389

Ciattaglia, Brunela (Estudiante, Escuela de Agrimensura. FCEIA. UNR) - ORCID: 0000-0002-7992-8287

Vicioso, Benito (Escuela de Agrimensura. FCEIA. UNR) - ORCID: 0000-0003-1809-0211

Peña, Hector (Escuela de Agrimensura. FCEIA. UNR) - ORCID: 0000-0003-2772-1054

Delorenzi, Dardo Javier (Escuela de Agrimensura. FCEIA. UNR) - ORCID: 0000-0003-3893-5648

Solís, Tania (Municipalidad de Rosario) - ORCID: 0000-0002-1738-3576

Figura 7. Identificadores ORCID a los autores de un conjunto de datos (RDA-UNR).

Fuente: Captura de pantalla <https://dataverse.unr.edu.ar/> (Balparda et al., 2023)

F4. Los (meta)datos están registrados o indexados en una fuente de búsqueda.

Los conjuntos de datos publicados con DOI en el Repositorio de Datos Académicos RDA-UNR son recolectados e indexados por DataCite Commons usando los metadatos DataCite. Asimismo, los metadatos están disponibles para otros servicios de búsqueda, incluyendo BASE, uno de los buscadores académicos más grandes a nivel mundial.

También, los metadatos de Schema.org están codificados en las páginas de inicio de los conjuntos de datos de Dataverse y desde allí son indexados por Google Dataset Search.

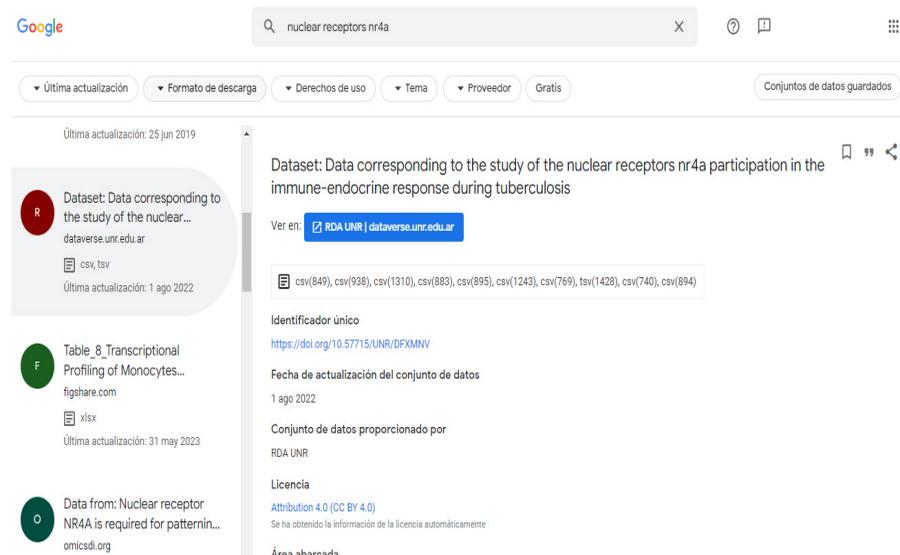


Figura 10. Conjunto de Datos del repositorio RDA-UNR en Google Dataset Search.

Fuente: <https://datasetsearch.research.google.com/>

A. Para ser Accesibles (Accesible)

A1. Los (meta)datos son recuperables mediante su identificador utilizando un protocolo de comunicación estandarizado.

Un protocolo de comunicación estandarizado consiste en un sistema de reglas que permite la transmisión de información entre sistemas de comunicación. De A1 se desprenden dos subprincipios:

- » A1.1: El protocolo debe ser abierto, gratuito y que se pueda implementar universalmente. En Dataverse, los datos y metadatos almacenados pueden ser accedidos a través de varios protocolos incluyendo el Protocolo de Transferencia de Hipertexto (HTTP), rsync sobre Secure Shell (SSH) y Representational State Transfer (REST) (2020) a través de la Interfaz de Programación de Aplicaciones (API) (2020). En el repositorio de la UNR, HTTP es el protocolo por defecto mediante el que los usuarios acceden a los (meta)datos. Si requieren acceso mediante otros protocolos lo deben gestionar con los curadores o administradores de sistemas del RDA-UNR.
- » A1.2: El protocolo permite un procedimiento de autenticación y autorización, cuando es necesario. Para el acceso a los datos a través de la API, Dataverse admite la autenticación basada tanto en sesiones como en claves de API, aunque en RDA-UNR no se han utilizado aún. En el software Dataverse se puede restringir el acceso a determinados archivos, en forma permanente o por un periodo de embargo. Hasta fines de 2023 todos los conjuntos de datos en RDA-UNR fueron de acceso abierto sin restricciones, de acuerdo con las políticas establecidas. En 2024 se presentó el primer caso de un archivo con restricción de acceso (<https://doi.org/10.57715/UNR/HLH331>).

A2. Los metadatos permanecen accesibles aun cuando los datos asociados no estén disponibles.

En circunstancias excepcionales, puede ocurrir que ciertos datos sean retirados del acceso público por motivos de fuerza mayor (por ejemplo, cuestiones legales). Aunque no se han registrado incidencias de datos retirados en RDA-UNR, la plataforma Dataverse tiene la capacidad de restringir el acceso a los archivos de un conjunto de datos, impidiendo su visualización o descarga, sin comprometer la disponibilidad de la información necesaria para citar dichos datos. En caso de restringir el acceso, es esencial brindar una explicación detallada en los metadatos sobre las causas por las que se retiran los datos.

I. Para ser Interoperables

I.1. Los (meta)datos utilizan un lenguaje formal, accesible, compartido y ampliamente aplicable para la representación del conocimiento.

Cuando se accede a la página principal de un conjunto de datos, esta contiene datos estructurados en un formato estándar y reconocible que programas informáticos pueden identificar y analizar automáticamente. Por ejemplo, los detalles de un conjunto de datos podrían estar marcados en la página HTML de manera que los motores de búsqueda y otras herramientas puedan extraer y comprender la información como el título del conjunto de datos, los autores, las fechas de publicación y otra información.

En el contexto de Dataverse, esto se logra mediante la adopción de Datos Enlazados con el soporte de JSON-LD para Schema.org. Esto implica que los metadatos de un conjunto de datos se presentan en un formato estandarizado que permite que la información sea indexable y combinable con otros datos que también utilicen este formato estándar en la web. Esta estructura facilita que las máquinas encuentren y combinen datos relacionados de diferentes fuentes, creando una representación más rica y completa del tema o entidad en cuestión.

Para poder ver claramente este proceso en el repositorio RDA-UNR se puede usar la herramienta F-UJI. Es un servicio web desarrollado para evaluar automáticamente la conformidad con los principios FAIR de objetos de datos de investigación a nivel de conjunto de datos basándose en las “Métricas de Evaluación de Objetos de Datos de FAIRsFAIR” (Devaraju et al., 2020).

Report:

Findable

FsF-F1-01D - Data is assigned a globally unique identifier. ✓ ↕

FAIR level: 3 of 3 advanced

Score: 1 of 1

Output:

```
{
  "guid": "https://doi.org/10.57715/UNR/EXIVRO",
  "guid_scheme": "doi"
}
```

Metric tests:

Test:	Test name:	Score:	Maturity:	Result:
FsF-F1-01D-1	Identifier is resolvable and follows a defined unique identifier syntax (IRI, URL)	1	3	✓
	test target: https://f-uji.net/vocab/identifier			
	condition: any of members of test target			
	tested on: https://f-uji.net/vocab/metadata/property/object_identifier			
	comment: identifier can be given as user input			
FsF-F1-01D-2	Identifier is not resolvable but follows an UUID or HASH type syntax	0		?
	test target: https://f-uji.net/vocab/identifier/unique			
	specifically: name-uuid, hash			

Figura 11. Fragmento de informe de la prueba automática de cumplimiento de principios FAIR en F-UJI. Interoperabilidad I1.

I2. Los (meta)datos usan vocabularios que siguen principios FAIR.

Otro aspecto de la interoperabilidad es el uso de vocabularios (meta)datos que siguen los principios FAIR.

Se pueden integrar en Dataverse vocabularios controlados conforme a FAIR y modelos de datos de manera manual, como podría ser la inclusión de palabras clave en la sección correspondiente de los metadatos, un proceso que requiere el apoyo y experiencia de curadores de datos especializados.

En RDA-UNR, al trabajar con curadores, en casi todos los conjuntos de datos se aplican vocabularios controlados como tesoro UNESCO, tesoro multilingüe DeCS/MeSH, MeSH, Tesoro UNBIS, AGROVOC Multilingual Thesaurus, entre otros.

Además, estos vocabularios FAIR pueden incorporarse en Dataverse mediante esquemas de metadatos a medida, presentándose al usuario como valores predefinidos o recomendados dentro de las plantillas de metadatos. Esto no se ha implementado en RDA-UNR. El software Dataverse aún no cuenta con soporte por defecto para vocabularios controlados y ontologías de mayor complejidad.

I3. Los (meta)datos incluyen referencias calificadas a otros (meta)datos.

Las referencias pueden agregarse en dos campos del esquema de metadatos generales en Dataverse, uno para “Publicación relacionada” (publicaciones que usan los datos del dataset) y otro para “Materiales relacionados” (otros objetos de investigación relacionados con el conjunto de datos). Actualmente, la información puede incluirse como texto libre en estos dos campos y la información no se exporta a DataCite. En una versión futura de Dataverse se mejorarán estas referencias.

Sin embargo, al hacer un testeo en la herramienta F-UJI se pueden reconocer las entidades relacionadas.

Interoperable

FsF-I1-01M - Metadata is represented using a formal knowledge representation language.

FAIR level: 3 of 3 advanced

Score: 2 of 2

Output:

```
[
  {
    "is_metadata_found": true,
    "serialization_format": "JSON-LD",
    "source": "structured_data"
  },
  {
    "is_metadata_found": true,
    "serialization_format": "JSON-LD",
    "source": "content_negotiate"
  }
]
```

Metric tests:	Test:	Test name:	Score:	Maturity:	Result:
	FsF-I1-01M-1	Parsable, structured metadata (JSON-LD, RDFa) is embedded in the landing page XHTML/HTML code	1	2	✔

test target: <http://f-uji.net/vocab/metadata/format>

specifically: name=RDF, JSON-LD, RDFa

condition: any of specified members of test target

Figura 12. Detalle de informe de prueba automática de cumplimiento de principios FAIR en F-UJI. Interoperabilidad I3.

Fuente: Captura de pantalla de <https://www.f-uji.net/index.php> con DOI <https://doi.org/10.57715/UNR/ULPO6K>

R. Para ser reutilizable (Reusable)

R1. Los (meta)datos disponen de atributos precisos y pertinentes. El principio se desarrolla con mayor detalle en tres subprincipios:

- » R1.1. Los (meta)datos se publican con una licencia de reutilización clara y accesible.

La información sobre la licencia de uso de los datos, así como la información sobre el acceso a los datos y los términos de uso, se incluyen por defecto en los metadatos de los datasets publicados en RDA-UNR.

La licencia por defecto de acuerdo con las políticas definidas en el RDA-UNR es CC-BY como se puede ver en la figura. Casi todos los conjuntos de datos usan esa licencia.

The screenshot shows a dataset page with the following details:

- Publicación relacionada:** Vanrell, S. R., Chelotti, J. O., Bugnon, L. A., Rufiner, H. L., Milone, D. H., Laca, E. A., & Galli, J. R. (2020). Audio recordings dataset of grazing jaw movements in dairy cattle. *Data in brief*, 30, 105623. doi: j.dib.2020.105623
- Notas:** Parameters for data collection: Cows grazing two forage species (pure alfalfa or pure tall fescue), each at two heights (tall and short) Description of data collection: Audios were recorded on three cows grazing on two forage species in individual grazing sessions: pure alfalfa or pure fescue microswards at two heights (tall or short)
- Acuerdo de licencia/uso de datos:** CC BY 4.0
- Condiciones del dataset:** Las Normas de nuestra Comunidad, así como también las buenas prácticas científicas, exigen que se otorgue el reconocimiento adecuado por medio de la citación. Utilice la cita de datos que se visualizará en la página del dataset.

Figura 13. Licencia visible en un conjunto de datos en RDA-UNR.

Fuente: Captura de pantalla de <https://dataverse.unr.edu.ar/>

La licencia de reutilización es correctamente recuperada por máquinas de acuerdo con la prueba de F-UJI.

The screenshot shows an automated test report for the F-UJI principle R1.1. The report includes the following information:

- FAIR level:** 3 of 3
- Score:** 2 of 2
- Output:**

```
[
  {
    "license": "http://creativecommons.org/licenses/by/4.0",
    "osi_approved": true,
    "details_url": "http://creativecommons.org/licenses/by/4.0"
  }
]
```
- Metric tests:**

Test:	Test name:	Score:	Maturity:	Result:
FsF-R1.1-01M-1	Licence information is given in an appropriate metadata element	1	1	✓
FsF-R1.1-01M-2	Recognized licence is valid (community specific or registered at SPDX)	1	3	✓
- Debug messages:**

Level:	Message:
INFO	Found CreativeCommons license -: http://creativecommons.org/licenses/by/4.0
SUCCESS	Found SPDX license representation (spdx url, osi_approved)
SUCCESS	Found licence information in metadata
INFO	Will consider all SPDX licenses as community specific licenses for FsF-R1.1-01M

Figura 14. Parte de informe de prueba automática de cumplimiento de principios FAIR en F-UJI. Reutilización R1.1.

Fuente: Captura de pantalla de <https://www.f-uji.net/index.php>

» R1.2. Los (meta)datos están asociados con su procedencia.

La segunda característica distintiva de los (meta)datos reutilizables, es que se les adjunta una información de procedencia exhaustiva. Esta procedencia abarca datos sobre el origen y la historia de los datos, como el modo en que fueron recolectados y procesados, así como los detalles sobre las personas o entidades involucradas en su gestión.

Dataverse ofrece metadatos enriquecidos, que engloban a los autores de los datos, colaboradores, entidades proveedoras y distribuidoras, además de datos relacionados, como aquellos datos que se han utilizado como insumo. Las modificaciones en los conjuntos de datos publicados quedan registradas automáticamente gracias al sistema de control de versiones de Dataverse. Asimismo, Dataverse facilita el registro de datos de procedencia de manera estructurada, acorde al modelo de datos PROV de la W3C, que establece un estándar para el intercambio de este tipo de información en la red (W3C). Por el momento no se usa una ontología específica para registrar los elementos de procedencia.

Versión del dataset	Resumen	Colaboradores	Publicado el
1.1	Metadatos de citas: Publicación relacionada (1 Añadido, 1 Modificado); Ver detalles	Dataverse Admin, Natalia Santucci	2023-11-07
1.0	Esta es la primera versión publicada.	Natalia Santucci, Dataverse Admin	2022-12-29

Figura 15. Información registrada de cambios y versiones en los conjuntos de datos RDA-UNR.

Fuente: Captura de pantalla de <https://dataverse.unr.edu.ar/>

» R1.3. Los (meta)datos cumplen con los estándares propios de una comunidad de un área de conocimiento determinada.

En los diferentes campos científicos existen múltiples estándares de (meta)datos. Dataverse soporta actualmente algunos de estos estándares. Además del esquema de metadatos para cita y recuperación, hay un esquema específico para las ciencias sociales, que cumple con partes de la Iniciativa de Documentación de Datos (DDI). También se ofrecen esquemas de metadatos menos detallados para datos de astronomía y astrofísica, así como para ciencias biológicas, metadatos geospaciales, metadatos de revistas, entre otros detallados en las Referencias de Metadatos de Dataverse.

En RDA-UNR, los esquemas de metadatos sí se utilizan en los conjuntos de datos de áreas específicas. Se requiere un intercambio entre investigadores y curadores para lograr que se completen estos campos más específicos. Estos intercambios se desarrollan en el marco de entrevistas de curación previamente pactadas.

Metadatos geospaciales	
Cobertura geográfica	Argentina, Santa Fe, Rosario
Unidad geográfica	Rosario
Metadatos de Ciencias Biológicas	
Tipo de factor	Extracto de molécula; Características genéticas; Tipo de tratamiento
Organismo	Otro
Otro organismo	Bacillus subtilis
Tipo de medición	perfil de expresión proteica; Otro
Otro tipo de medición	Cuenta de células; Hydrogen/Deuterium exchange; anisotropy; fatty acid percentage
Tipo de tecnología	electroforesis en gel; espectrometría de masas; secuenciación de nucleótidos; Otro
Plataforma tecnológica	Agilent; Otro
Otra plataforma tecnológica	Cromatógrafo gaseoso Agilent modelo 7890B acoplado a un Espectrómetro de Masas Agilent modelo 5977A

Figura 16. Detalle de los metadatos de Ciencias Biológicas en un conjunto de datos en RDA-UNR.

Fuente: Captura de pantalla de <https://dataverse.unr.edu.ar/dataset.xhtml?persistentId=doi:10.57715/UNR/PTDCEY>

La comunidad de usuarios de Dataverse está trabajando en establecer esquemas de metadatos, incluyendo vocabularios controlados, para más áreas de investigación. Además de los esquemas de metadatos que Dataverse ofrece por defecto, los estándares comunitarios relevantes para un dominio pueden implementarse creando bloques de metadatos personalizados. En el caso de RDA-UNR no se han agregado nuevos campos de metadatos, pero sí se especifica cómo se deben completar algunos campos, por ejemplo, la Descripción (Abstract).

En relación con la gestión de archivos, Dataverse puede extraer automáticamente metadatos de los archivos FITS, un formato estándar en astronomía. En RDA-UNR hasta fines de 2024, no se subieron este tipo de archivos por lo que esta funcionalidad no se ha puesto en práctica.

Una funcionalidad importante de Dataverse, para archivos que contienen datos en forma de tablas como los creados en R, Stata o Excel, es que el software realiza una conversión automática a un formato de texto simple separado por tabulaciones, considerado más robusto para la preservación de datos a largo plazo. Durante este proceso, también se extraen los metadatos descriptivos de cada variable disponible en las columnas de dichos archivos.

Aunque el equipo de curadores del RDA-UNR sugiere a los investigadores que, además de los datos tabulares en el formato original (por ejemplo, SPSS) se adjunten en formatos abiertos (.csv), también se usa la funcionalidad por defecto de Dataverse como puede verse en la siguiente figura:

Figura 17. Acceso a los archivos de datos tabulares en RDA-UNR.



Fuente: Captura de pantalla de <https://dataverse.unr.edu.ar/dataset.xhtml?persistentId=doi:10.57715/UNR/OJ4BUD>

En resumen, en cuanto al cumplimiento de los principios FAIR se trabaja en todos sus elementos, en algunos con mayor cumplimiento que otros en el análisis automático en fase de experimentación de la herramienta F-UJI.

Summary:



Figura 18. Un ejemplo de informe de cumplimiento de principios FAIR en F-UJI.

Fuente: Captura de pantalla de <https://www.f-ujl.net/index.php>

En RDA-UNR se atienden otros aspectos relativos a la reutilización de los datos que no se pueden verificar con análisis automáticos y que se dan en el proceso de curación de datos, un trabajo cualitativo e interdisciplinario entre curadores e investigadores. Este proceso se detalla en el próximo apartado.

El proceso interdisciplinario de curación de datos

El equipo del RDA-UNR ofrece acompañamiento en la curación de los datos, su organización y descripción, de modo que otros usuarios puedan comprender la estructura y el contenido del conjunto de datos y así los puedan citar y reutilizar.

En los repositorios de datos, la curación de datos de investigación implica la gestión activa y la valoración de los datos a lo largo de su ciclo de vida para hacerlos útiles y accesibles para futuros usos, es decir, su reutilización (Boté-Vericad et al., 2023).

La curación de datos de investigación no es simplemente un proceso técnico, sino que está imbuido de negociaciones, construcción social entre los investigadores y curadores, aspectos retóricos y contextuales, lo cual es coherente con las ideas presentadas por Knorr-Cetina (2005) sobre la ciencia como una práctica constructivista y contextual.

En la primera consulta o entrevista de curación, muchas veces se sugiere a los investigadores que modelen el flujo de la investigación, incluyéndose los pasos y procesos donde se generan datos. Este proceso no solo identifica la procedencia de los datos, sino que también ayuda a los curadores a entender la estructura y la metodología de la investigación, como se puede ver en la figura 19. El flujo de investigación y datos es especialmente útil en datos experimentales (Esteva et al., 2019).

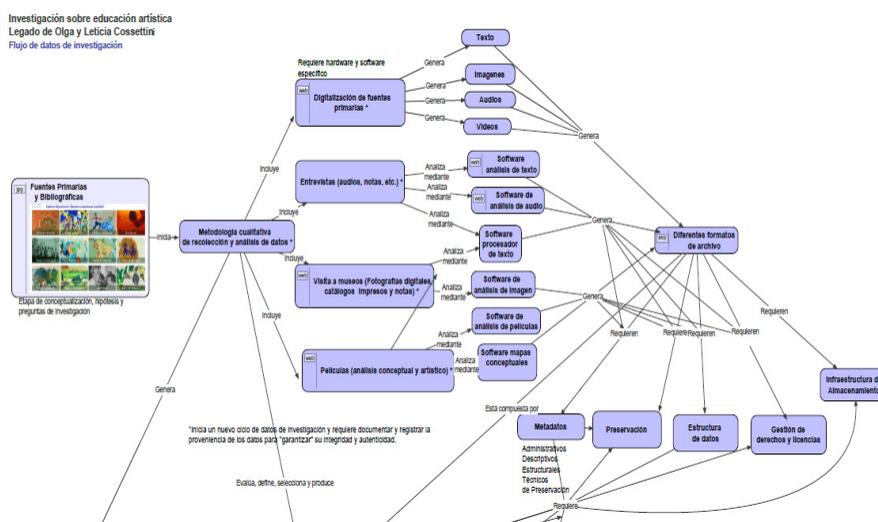


Figura 19. Ejemplo de un flujo de investigación y sus pasos.

Fuente: Bongiovani, 2024.

Este modelado del flujo de la investigación también ayuda a los investigadores a decidir qué datos van a compartir. Esto puede no ser tan claro. Una investigadora en ciencias médicas pregunta en una capacitación (UNR, 2022): “si haces una tomografía tenés 500 mil cortes y 500 mil estudios mediante los cuales uno saca una conclusión. No se pueden subir los 500 mil cortes, entonces, pregunto: ¿cómo se hace? ¿qué se sube?”. Es clave dilucidar qué datos son la evidencia para esa investigación en particular, cuál es el valor de esos datos, qué consentimiento informado existió para compartir o no imágenes de pacientes.

Otro elemento que brinda confianza a los investigadores es la ausencia de una prescripción con respecto a los datos que se esperan. Son los propios investigadores quienes definen qué datos (objetos) serán compartidos y pueden utilizarse como evidencia para respaldar sus afirmaciones (Leonelli, 2015), haciendo referencia a situaciones de investigación concretas.

En la consulta de curación los profesionales de la información preguntan sobre los siguientes aspectos: motivaciones para generar u obtener los datos (objetivos de la investigación), tipos de datos y tipos de archivos, unidad de análisis/temporalidad y cobertura geográfica, unidades/conjunto de datos dentro del proyecto, cantidad de archivos por dataset, métodos para la descripción (metadatos estandarizados o no, diccionario de datos), autores, instrumentos utilizados, análisis posteriores realizados, aspectos legales o de propiedad de los datos, consentimientos informados, entre otros elementos. En este nivel, los profesionales de la información aportan su experiencia en el tratamiento sistemático de la información, poniendo en juego sus competencias en metadatos estandarizados, clasificación, representación temática y organización.

Los curadores motivan a los investigadores para que clarifiquen estos aspectos a fin de completar toda la información en los campos de metadatos disponibles en el Dataverse RDA-UNR o en los archivos que acompañan la investigación. Al ser un repositorio multidisciplinario, se presenta una gran variedad de proyectos de investigación y tipos de datos.

Se explica a los investigadores que, para que los datos se puedan reutilizar, es necesario documentar la investigación. La documentación de los datos informa cómo y por qué se realizaron ciertos pasos o se tomaron ciertas decisiones, dado que lo que es obvio para los investigadores que generaron los datos, puede no ser intuitivo

para otros investigadores o incluso para ellos mismos en el futuro. Documentar estas decisiones puede ayudar en la reutilización, ya sea replicación, reproducción o de otro tipo, y la validación de los datos (Bongiovani, 2023a).

Como ya se indicó en la sección anterior, la reutilización de datos implica no solo acceder a los datos, sino también poder aplicarlos en nuevos contextos o combinarlos con otros conjuntos de datos para generar nuevo conocimiento. El ejercicio en la curación es ponerse en el lugar de otro y anticipar qué documentación es necesario incluir para que se puedan reutilizar.

La mayoría de las veces los investigadores no tienen claro qué son los metadatos, esto fue evidenciado en las capacitaciones y consultas de curación, por lo cual se les dan ejemplos concretos en el repositorio, además de estar disponible la definición en el Glosario. También se explica qué se espera de la descripción. Por ejemplo, el curador puede explicar que el título que se carga en el repositorio no es el título del artículo publicado. Es importante recordar que toda la descripción se centra en el conjunto de datos que se va a subir, no sobre el artículo. Se debería incluir un título claro para un usuario externo que quiera reutilizar los datos y venga desde un buscador web.

Se trata de un proceso de negociación porque, puede pasar que el investigador decida de todos modos poner el título del artículo o que acepte el consejo de los curadores. Estas decisiones pueden tener más que ver con el tiempo que los investigadores tengan disponible para dedicar a la curación de los datos. Además, los investigadores responden a las normativas de las revistas de publicación, que de manera creciente están solicitando la apertura y publicación de los datos al momento de enviar los manuscritos.

Un elemento esencial, muchas veces trabajoso, en el proceso de descripción de los datos es la escritura del resumen de datos (abstract) porque ayuda a que los usuarios potenciales puedan descubrirlos, entender qué son, qué alcance tienen, para qué le pueden servir los datos compartidos y decidir reutilizarlos. En la experiencia de trabajo con los primeros conjuntos de datos se hizo evidente que era difícil lograr esa descripción pensada en la reutilización.

Por eso, además de las capacitaciones y consultas, el RDA-UNR puso a disposición una guía para ayudar a elaborar esa correcta descripción basada en Smith y Phillips (2021). En esta guía con preguntas, se ofrecen consejos prácticos para la escritura, pero en general, es un aspecto que requiere varias idas y vueltas entre investigadores y curadores para lograr claridad y precisión en la descripción. Un aspecto de especial interés para el RDA UNR, es que los investigadores destaquen la relevancia de sus datos en un lenguaje accesible para públicos amplios, de modo tal que usuarios diversos puedan comprender la significancia del aporte de la ciencia financiada por el sector público. Este aspecto, más cercano a la divulgación de la ciencia, se refuerza con las piezas de comunicación que se han denominado “Historias de datos” y que se detallan más adelante.

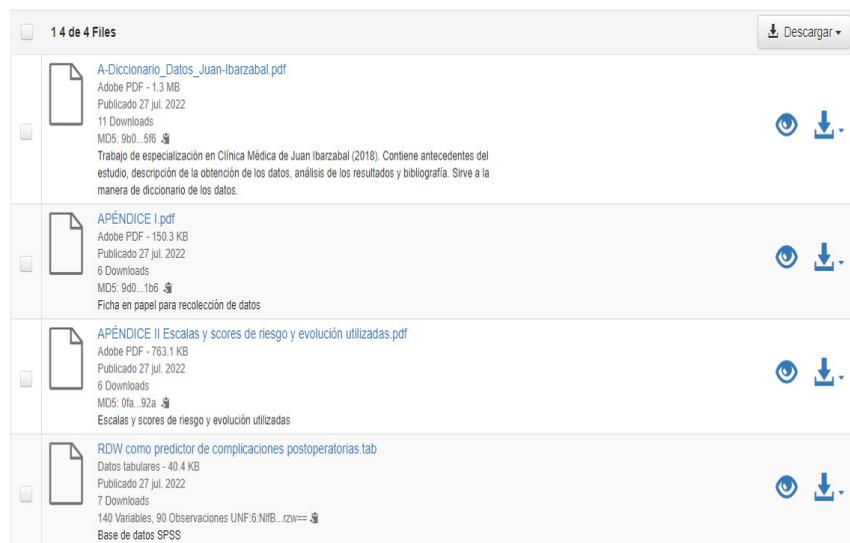
Dependiendo de la necesidad de los investigadores, los curadores muestran cómo se cargan los metadatos y datos en el repositorio y se deja que ellos completen los campos de metadatos hasta donde les sea posible, siempre aclarando que es un proceso incremental. Es decir, primero se pueden agregar los metadatos de cita y, una vez que se guarda esta información, aparecerán otros campos más específicos de la disciplina. Se aclara que es importante que se complete todo lo que se aplique a ese conjunto de datos en particular.

En ese proceso de demostración práctica, se aborda la importancia de los identificadores persistentes (Bongiovani et al., 2023b). Se solicita que los autores tengan su ORCID. Esta información también está disponible en forma de infografía “¿Qué son los identificadores persistentes?”.

En Dataverse se documentan los datos a nivel del proyecto: incluye información sobre el estudio realizado, cuáles fueron las preguntas de investigación, el tipo de datos que se recopilaron para responder a estas preguntas, metodología, tecnología, software, periodos de tiempo, y otros campos específicos de acuerdo con la disciplina como unidad de análisis, universo, tipo de instrumento, tipo de diseño, etc. Este paso lo realizan los investigadores y muchas veces lo completan los curadores consultando a los autores.

En RDA-UNR también se pide a los investigadores una documentación a nivel de los datos: incluye información sobre los archivos de datos individuales, como la transcripción de una entrevista, las variables particulares (diccionario de datos, libro de códigos, etc.) junto con los datos. Los curadores abren los archivos y revisan su contenido para verificar que sean completos y comprensibles (nivel de curación 3 de acuerdo con Digital Curation Network).

Figura 20. Ejemplo de un diccionario de datos en RDA-UNR.



Fuente: Captura de pantalla de Ibarzabal et al., 2022 (RDA-UNR) <https://doi.org/10.57715/UNR/OJ4BUD>

Dependiendo del tipo de estudio, por ejemplo, en una encuesta, se deben incluir también elementos como cuestionarios en blanco, formularios de consentimiento informado que hayan firmado los participantes del estudio, y si son registros audiovisuales, listas de datos que detallan fechas, lugares y otros metadatos clave.

Otra recomendación de los curadores es organizar los archivos de datos en carpetas, especialmente cuando se manejan grandes volúmenes de información o se busca reflejar la estructura de la investigación. Estas carpetas funcionan como categorías que agrupan los archivos relacionados, facilitando su navegación y comprensión.

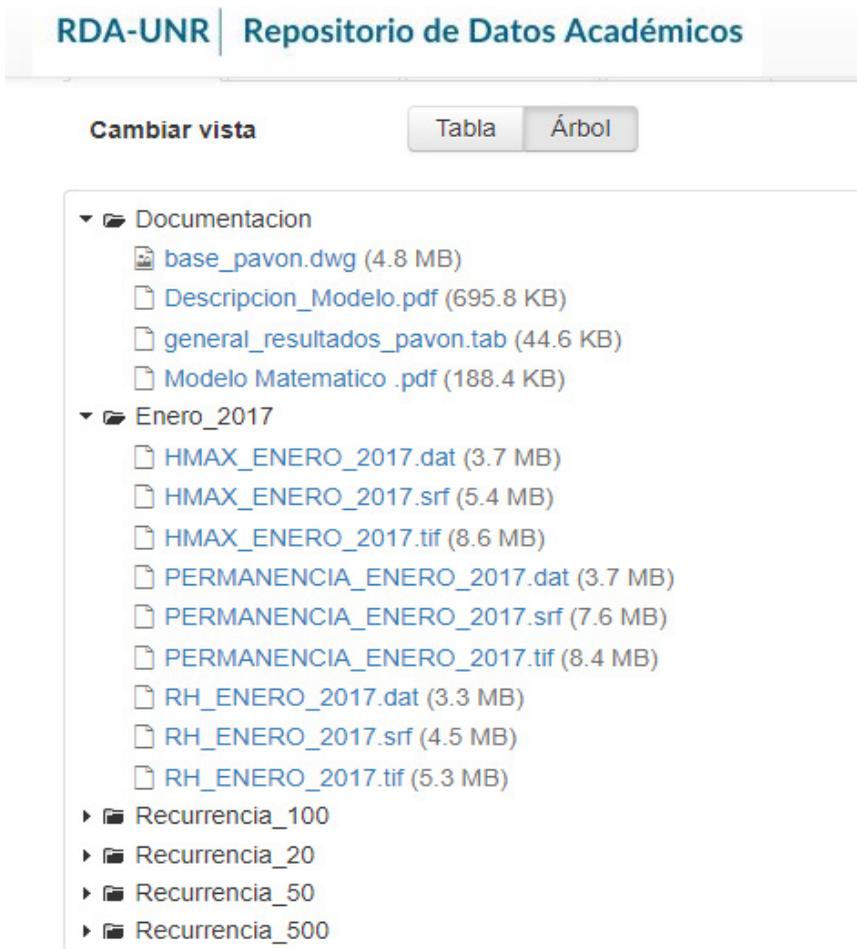


Figura 21. Organización de los archivos en carpetas en RDA-UNR.

Fuente: Captura de pantalla de Stenta et al., 2022. <https://doi.org/10.57715/UNR/THCoKS>

A partir de los conjuntos de datos publicados en 2023, se solicita también a los investigadores la inclusión de un archivo README, considerado una parte importante de la documentación de un conjunto de datos. Este archivo, diseñado para ser legible por humanos, suele contener información clave que facilita la comprensión y el uso de los datos por parte de otros usuarios. En la tabla se presenta un ejemplo del tipo de contenidos que puede incluir. Actualmente, el equipo de curadores del RDA-UNR está desarrollando una plantilla específica para este propósito.

Conclusiones y desafíos a futuro

El sistema de investigación y desarrollo se enmarca en el paradigma de la ciencia abierta; potentes herramientas computacionales, entre las que se encuentra la inteligencia artificial generativa, favorecen el desarrollo de un ecosistema de datos masivos que los investigadores pueden aprovechar para crear nuevo conocimiento. El valor de los repositorios de datos no es solo presente, sino una inversión a largo plazo (Borgman, Scharnhorst y Golshan, 2019). A pesar del costo que implican, el hecho de ser intensivos en mano de obra y que es difícil medir su impacto, deben demostrar constantemente su valor a sus comunidades, y ser reconocido por

políticas, mandatos institucionales y leyes. Argentina y la UNR han sido pioneras en el reconocimiento de este fenómeno, promoviendo políticas, infraestructuras y formando recursos humanos para garantizar la integridad y disponibilidad de los datos. Todas estas acciones hicieron posible la creación del RDA UNR, y la colaboración entre investigadores y diversos actores dentro de la institución, destacándose el rol de los profesionales de la información.

Por otra parte, para mejorar las capacidades se está trabajando en ampliar los curadores especializados a más disciplinas en la medida en que más bibliotecarios de la Universidad se van capacitando e involucrando en las prácticas de gestión de datos de investigación. Esto permitirá completar el equipo de trabajo con curadores de todas las facultades, y la conformación del comité asesor del RDA UNR. Se requiere multiplicar las acciones formativas y de alfabetización para llegar a los distintos grupos de investigación relacionados con cada una de las unidades académicas.

Se está avanzando en la creación de una estrategia de “líderes de datos”, similar a las iniciativas de campeones de datos (Data Champions) que se desarrollaron en algunas universidades europeas, por ejemplo, en la Universidad de Cambridge y la Universidad de Tecnología de Delft (TU Delft) (Clare et al., 2019). El propósito es contar con referentes dentro de la comunidad académica que ejemplifiquen y fomenten buenas prácticas en la gestión de datos y la ciencia abierta. Estos líderes de datos sirven como referentes y promotores en la gestión de datos, ofreciendo sus experiencias, herramientas y recomendaciones a sus colegas para inspirar y guiar en este campo. Este proceso se está dando naturalmente en algunas disciplinas.

Integrar los resultados del RDA UNR al Observatorio de Ciencia Abierta de la UNR que está en formación, es uno de los avances a concretar en el corto plazo.

Una de las necesidades detectadas en las comunidades científicas con las que se trabaja, es avanzar en un entendimiento con otras instituciones de investigación a nivel nacional para coordinar y simplificar los procesos en cuanto a apertura de datos de investigación, a fin de facilitar y favorecer la participación de los investigadores en el proceso.

En futuras publicaciones, se detallarán las acciones implementadas en el marco del Plan de Comunicación del Repositorio de Datos Académicos (RDA-UNR). Este análisis proporcionará una visión completa de las estrategias diseñadas para fortalecer la difusión y accesibilidad de los datos de investigación. Se espera que estos esfuerzos no solo amplíen el alcance de la comunidad académica, sino también promuevan el acceso abierto y el uso efectivo de los recursos científicos disponibles.

Agradecimientos

A todo el equipo de trabajo y curadores del Repositorio de Datos Académicos RDA-UNR: Analía Salazar, Dolores Quintana, Agustín Alfieri, Gisela Chiappero, Ana Paula Gutiérrez, Laura Balparda, Diego López y María Esteva. A todos las investigadoras e investigadores de la UNR que trabajan cada día para brindar sus datos de investigación de manera generosa al mundo.

Referencias bibliográficas

- » Abadal, Ernest y Lluís Anglada. 2020. Ciencia abierta: cómo han evolucionado la denominación y el concepto. En *Anales de documentación*. Vol. 23, no. 1. <<https://doi.org/10.6018/analesdoc.378171>>
- » Anglada, Lluís y Ernest Abadal. 2018. ¿Qué es la ciencia abierta? En *Anuario ThinkEPI*. Vol. 12, 292-298. <<https://doi.org/10.3145/thinkepi.2018.43>>
- » Balparda, Laura Rita, Héctor Del Valle, Diego Alejandro Germán López, María Cecilia Torralba, Florencia Tazzioli, Brunela Ciattaglia, Benito Vicioso, Hector Peña, Dardo Javier Delorenzi y Tania Solís. 2023. Datos de: Huella Urbana de la Ciudad de Rosario, Santa Fe, Argentina [Data set]. RDA UNR. <<https://doi.org/10.57715/UNR/EXIVRO>>
- » Bongiovani, Paola Carolina, Griselda Guarnieri, Dominique Babini y Fernando Ariel López. 2014a. Acceso abierto en la Universidad Nacional de Rosario: Necesidades y prácticas de los docentes/investigadores. En *Información, cultura y sociedad*. No. 30, 13-33. <http://www.scielo.org.ar/scielo.php?script=sci_arttext&pid=S1851-17402014000100002&lng=es&tlng=es>
- » Bongiovani, Paola Carolina y Luis Martínez-Urbe. 2014b. Necesidades de gestión de datos científicos en Argentina. El caso de la Universidad Nacional de Rosario. En *Anais das sessões temáticas e pôsters. Conferência Internacional Acesso Aberto, Preservação Digital, interoperabilidade, Visibilidade e Dados Científicos*. Biredial Istec 2014. <<http://hdl.handle.net/10183/108316>>
- » Bongiovani, Paola Carolina y Sandra E. Miguel. 2019a. ¿Cuán abierta es la producción científica de los investigadores argentinos de Ciencias Sociales? En *Palabra clave*. Vol. 9, no. 1, 80-80. <<http://dx.doi.org/https://doi.org/10.24215/18539912e080>>
- » Bongiovani, Paola Carolina, Claudia Voras y Alejandro Pérez Bigot. 2019b. La institucionalización del acceso abierto en la Universidad Nacional de Rosario. En *La cultura de los datos: Actas del II Congreso Internacional de la Asociación Argentina de Humanidades Digitales (2018 : Rosario)*. La Plata: Universidad Nacional de La Plata. Facultad de Humanidades y Ciencias de la Educación; Rosario: Universidad Nacional de Rosario. (Trabajos, comunicaciones y conferencias; 42). <<https://www.libros.fahce.unlp.edu.ar/index.php/libros/catalog/book/153>> [Consulta: 20 marzo 2025]
- » Bongiovani, Paola Carolina, María Esteva, Fernando Díaz Pacífico, Ana Paula Gutierrez, Laura Rita Balparda, Diego Mestre, Analía Salazar y Paulina Freán. 2022. Planificación e Implementación del Repositorio de Datos Académicos de la Universidad Nacional de Rosario (RDA-UNR): una metodología colaborativa e interdisciplinaria. En *Conferencia Internacional BIREDIAL ISTEAC*. RepHipUNR <<http://hdl.handle.net/2133/24576>>
- » Bongiovani, Paola Carolina. 2023a. Capacitación Módulo 2 Trae tus datos. Cómo publicar Datos de investigación en el Repositorio de Datos Académicos RDA-UNR. RepHipUNR <<https://rehip.unr.edu.ar/handle/2133/26503>> [Consulta: 20 marzo 2025]
- » Bongiovani, Paola Carolina, Analía Salazar y Paulina Freán. 2023b. Implementación de PIDs en América Latina. Repositorio de Datos Académicos RDA-UNR dataverse.unr.edu.ar. Zenodo. <<https://doi.org/10.5281/zenodo.7860470>>

- » Bongiovani, Paola Carolina, Analía Salazar, Paulina Freán, Fernando Díaz Pacífico, Ana Paula Gutierrez, Dolores Quintana, Laura Balparda, Diego López, Hernán Stenta y Natalia Santucci. 2023c. Trabajo en equipo y compromiso, claves para la implementación exitosa del Repositorio de Datos Académicos RDA-UNR. Semana de Ciencia Abierta LA Referencia 2023. [Video]. YouTube. <https://youtu.be/g98WiVT9nlro?si=l4MHdFvfgvQGcNpw>
- » Bongiovani, Paola Carolina. 2024. Factores individuales e institucionales que influyen en la intención de compartir datos de investigación en repositorios. El caso de la Universidad Nacional de Rosario. Madrid: Universidad Carlos III de Madrid. Tesis doctoral. <<https://hdl.handle.net/10016/44153>>
- » Borgman, Christine L., Andrea Scharnhorst y Milena S. Golshan. 2019. Digital data archives as knowledge infrastructures: Mediating data sharing and reuse. En *Journal of the Association for Information Science and Technology*. Vol. 70, no. 8, 888-904. <<https://doi.org/10.1002/asi.24172>>
- » Boté-Vericad, Juan-José, Ana Carballo-García, Mònica Bautista-Villaescusa y Sharon C. Healy. 2023. Research data repositories in the RDM cycle: challenges and strengths for curators/data stewards. En *AIBstudi*. Vol. 63, no. 3. <<https://doi.org/10.2426/aibstudi-13892>>
- » Brochu, Lauren y Jane Burns. 2019. Librarians and Research Data Management—A Literature Review: Commentary from a Senior Professional and a New Professional Librarian. En *New Review of Academic Librarianship*. Vol. 25, no. 1, 49-58. <<https://doi.org/10.1080/13614533.2018.1501715>>
- » Conzett, Philipp. 2020. DataverseNO: A national, generic repository and its contribution to the increased FAIRness of data from the long tail of research. En *Ravnetrykk*. No. 39. <<https://doi.org/10.7557/15.5514>>
- » Crosas, Mercè. 2011. The Dataverse Network®: An Open-Source Application for Sharing, Discovering and Preserving Data. En *D-Lib Magazine*. Vol.17, no. 1/2. <<https://doi.org/10.1045/january2011-crosas>>
- » Crosas, Mercè. 2020. Fair Principles and Beyond: Implementation in Dataverse. En *Septentrio Conference Series*. No. 2. <<https://doi.org/10.7557/5.5334>>
- » Dataverse Project. s.f. Dataverse. <<https://dataverse.org/>> [Consulta: 20 marzo 2025]
- » Devaraju, Anusuriya, Robert Huber, Mustapha Mokrane, Patricia Herterich, Linas Cepinskas, Jerry de Vries, Herve L'Hours, Joy Davidson y Angus White. 2020. FAIRsFAIR Data Object Assessment Metrics (o.4). Zenodo <<https://doi.org/10.5281/zenodo.4081213>>
- » Esteva, Maria, Craig Jansen, Pedro Arduino, Mahyar Sharifi-Mood, Clint N. Dawson y Josue Balandrano-Coronel. 2019. Curation and publication of simulation data in DesignSafe, a natural hazard engineering open platform and repository. En *Publications*. Vol. 7, no. 3, 51. <<https://doi.org/10.3390/publications7030051>>
- » Force11. 2020. Guiding Principles for Findable, Accessible, Interoperable and Re-Usable Data Publishing Version B1.o. <<https://force11.org/info/guiding-principles-for-findable-accessible-interoperable-and-re-usable-data-publishing-version-b1-o/>> [Consulta: 20 marzo 2025]
- » Hernández-Pérez, Tony y María-Antonia García-Moreno. 2013. Datos abiertos y repositorios de datos: nuevo reto para los bibliotecarios. En *Profesional de la información*. Vol. 22, no. 3, 259-263. <<https://doi.org/10.3145/epi.2013.may.10>>

- » Ibarzabal, Juan Ignacio, Mariana Lagrutta, Juan Carlos Pendino y Roberto Leandro Parodi. 2022. Dataset Relación entre el Aumento de la Amplitud de Distribución del Diámetro Eritrocitario (Red Cell Distribution Width o RDW) y la morbilidad postquirúrgica. [Data set]. RDA UNR <<https://doi.org/10.57715/UNR/OJ4BUD>>
- » Knorr-Cetina, Karin. 2005. La fabricación del conocimiento. Un ensayo sobre el carácter constructivista y contextual de la ciencia. Buenos Aires: Universidad Nacional de Quilmes Editorial.
- » Leonelli, Sabina. 2015. What counts as scientific data? A relational framework. En *Philosophy of Science*. Vol. 82, no. 5, 810-821. <<https://doi.org/10.1086/684083>>
- » Ley 26.899. 2013. Creación de Repositorios Digitales Institucionales de Acceso Abierto, Propios o Compartidos. <https://repositoriosdigitales.mincyt.gob.ar/files/Boletin_Oficial_Ley_26899.pdf> [Consulta: 20 marzo 2025]
- » Méndez Rodríguez, Eva María. 2021. Open Science por defecto: La nueva normalidad para la investigación. En *Arbor: Ciencia, pensamiento y cultura*. Vol. 197, no. 799. <<https://doi.org/10.3989/arbor.2021.799002>>
- » Ministerio de Ciencia, Tecnología e Innovación de la Nación. 2021. Relevamiento de Actividades Científicas y Tecnológicas (RACT).
- » National Institutes of Health NIH. 2020. Final NIH Policy for Data Management and Sharing (NOT-OD-21-013). Bethesda: National Institutes of Health. <<https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html>> [Consulta: 20 marzo 2025]
- » Pinfield, Stephen, Andrew M. Cox y Jen Smith. 2014. Research data management and libraries: Relationships, activities, drivers and influences. En *PLoS ONE*. Vol. 9, no. 12, 1-28. <<https://doi.org/10.1371/journal.pone.0114734>>
- » Reglamento de la Ley 26.899. 2016. <https://repositoriosdigitales.mincyt.gob.ar/files/Boletin_Oficial_Resolucion_753.pdf> [Consulta: 20 marzo 2025]
- » Salazar, Analía, Paulina Freán, Gisela Chiappero, Dolores Quintana, Paola Carolina Bongiovani y Agustín Alfieri. 2023. Implementación de Identificadores Persistentes (PID) para aumentar el impacto del Repositorio De Datos Académicos RDA-UNR (dataverse.unr.edu.ar). En XVII 385 Jornadas de Ciencia, Tecnología e Innovación. Rosario: Universidad Nacional de Rosario. <<https://jornadasctei.unr.edu.ar/implementacion-de-identificadores-persistentes-pid-para-aumentar-el-impacto-del-repositorio-de-datos-academicos-rda-unr-dataverse-unr-eduar/>> [Consulta: 20 marzo 2025]
- » San Martín, Patricia Silvana, Griselda Guarnieri, Guillermo Luján Rodríguez, Paola Carolina Bongiovani y Alejandro Roberto Sartorio. 2010. El dispositivo hipermedial dinámico Campus virtual UNR. *RepHipUNR* <<http://hdl.handle.net/2133/1390>>
- » Smith, Michael y Dan Phillips. 2021. How To Write A (Good) Data Description: Developing Best Practice. Research Data Alliance Virtual Plenary 17 (RDA), Edinburgh (virtual). Zenodo. <<https://doi.org/10.5281/zenodo.4709835>>
- » Stenta, Hernán, Gerardo Riccardi, Pedro Basile y Carlos Scuderi. 2022. Datos de: Modelación matemática hidrológica-hidráulica del escurrimiento superficial en la cuenca del A° Pavón (Santa Fe, Argentina) [Data set]. RDA UNR. <<https://doi.org/10.57715/UNR/THCoKS>>
- » Universidad Nacional de Rosario. s.f. Datos abiertos. Repositorio de Datos Académicos UNR. <<https://dataverse-info.unr.edu.ar/sobre-datos-abiertos/#datos>> [Consulta: 20 marzo 2025]

- » Universidad Nacional de Rosario. 2016. Resolución de Consejo Superior N° 1842/2016. RepHipUNR <<http://hdl.handle.net/2133/12981>> [Consulta: 20 marzo 2025]
- » Universidad Nacional de Rosario. 2017. Resolución N° 5425/2017. RepHipUNR <<http://hdl.handle.net/2133/12982>>
- » Universidad Nacional de Rosario. 2018. Resolución N° 2159/2018. RepHipUNR <<http://hdl.handle.net/2133/1298>>
- » Vicente-Saez, Ruben y Clara Martínez-Fuentes. 2018. Open Science now: A systematic literature review for an integrated definition. En *Journal of business research*. Vol. 88, 428-436. <<https://doi.org/10.1016/j.jbusres.2017.12.043>>
- » Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak y Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. En *Scientific data*. Vol. 3, no. 1, 1-9. <<https://doi.org/10.1038/sdata.2016.18>>