CONFERENCIA INTERNACIONAL
BIREDIAL-ISTEC '24
22-24 de octubre de 2024 SANTIAGO•CHILE

¿En quién pienso cuando comparto mis datos de investigación?

Eje temático
Datos abiertos

Palabras claves
Dataverse, datos abiertos, calidad, procesamiento de la información, repositorio de datos

Dataverse, open data, quality, information processing, data repository

María Hidalgo Gutiérrez. Universidad de Costa Rica, Vicerrectoría de Investigación, *maria.hidalgogutierrez@ucr.ac.cr*
0009-0006-4716-2551

Meilyn Garro Acón. Universidad de Costa Rica, Vicerrectoría de Investigación, *meilyn.garro@ucr.ac.cr*
0000-0002-8400-9891

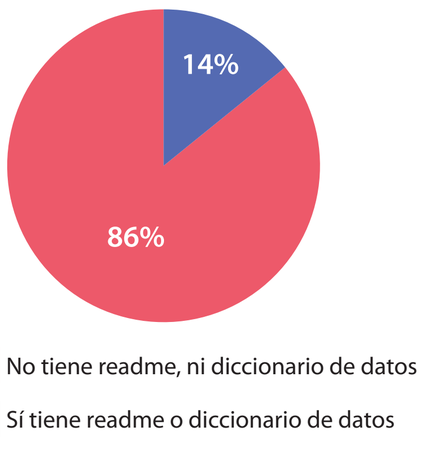
La implementación de espacios para gestionar datos de investigación es una necesidad desde hace más de 10 años (Hernández-Pérez & García-Moreno, 2013), y Dataverse es una de las herramientas empleadas para almacenar los datos de investigación en instituciones de educación superior en América Latina. La implementación de Dataverse se complementa con la definición de características deseables de los sets de datos, por ejemplo, los principios FAIR (FAIR Principles, 2016), así como características deseables para los repositorios de datos (Morales Vargas & Codina, 2019). A pesar de los esfuerzos por normalizar la descripción de los sets de datos y mejorar sus características, existe una brecha entre las recomendaciones y la realidad que limita las posibilidades de los potenciales usuarios para reutilizar los sets de datos. El objetivo de este trabajo es dimensionar la posibilidad de reutilizar los sets de datos en los Dataverse de instituciones de educación superior en Latinoamérica con base en criterios técnicos.

Metodología

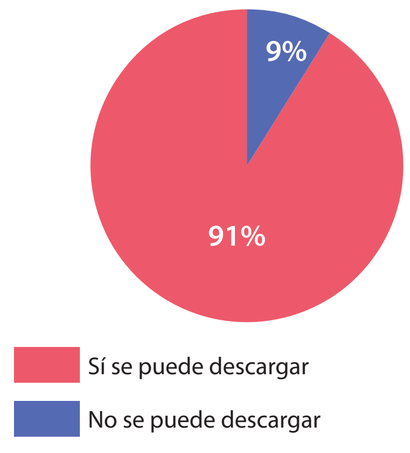
Se hizo una selección aleatoria de 4 repositorios institucionales repositorios de instituciones de educación superior listados en Dataverse Installations Around the World (dataverse.org/installations): **1)** Repositorio de Datos de Investigación Universidad del Rosario; **2)** Portal de datos abiertos, Pontificia Universidad Católica del Perú; **3)** Repositorio de Datos Académicos RDA-UNR; y **4)** Repositorio de Datos de Investigación de la Universidad Nacional de La Plata. Se definió un muestreo simple al azar para cada uno de los Dataverse seleccionados, con un error máximo del 10% y un nivel de confianza del 95%. En total, se revisaron 121 ítems.

Criterios evaluados:

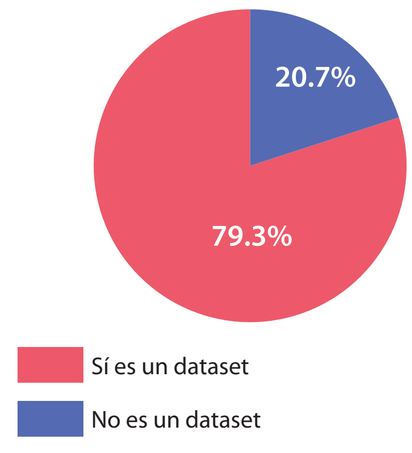
- **Descripción del conjunto de datos**: Presencia de readme o diccionario, descripción completa en los metadatos. La descripción debe ser suficiente para entender y reutilizar el set de datos.



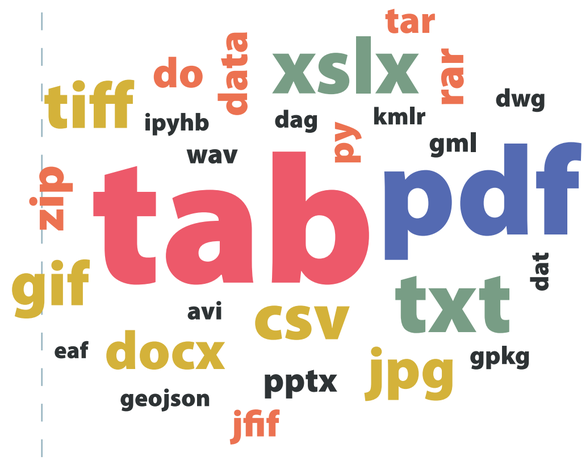
- **Conjunto de datos abierto**: Si el set de datos y demás archivos se pueden descargar.



- **Coherencia de la clasificación**: Tipo de documento clasificado como dataset es realmente un data set. Por ejemplo, que diga que es set de datos pero es un Plan de Gestión de Datos (PGD).



- **Formato del conjunto de datos**: lista de formatos incluidos en los sets de datos.



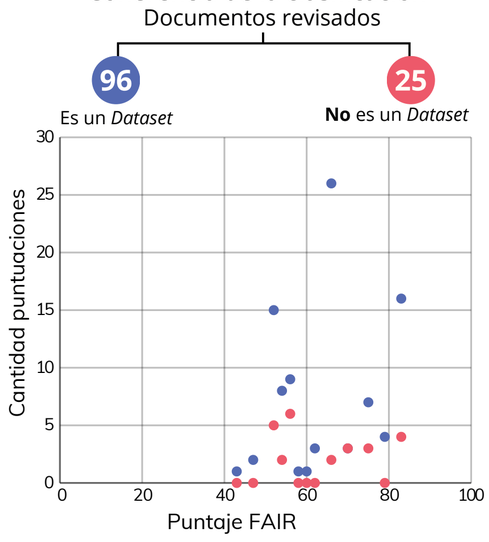
- **Relación entre la calificación FAIR y coherencia de clasificación:**

1- Relación entre documentos mal clasificados con puntaciones en FAIR.

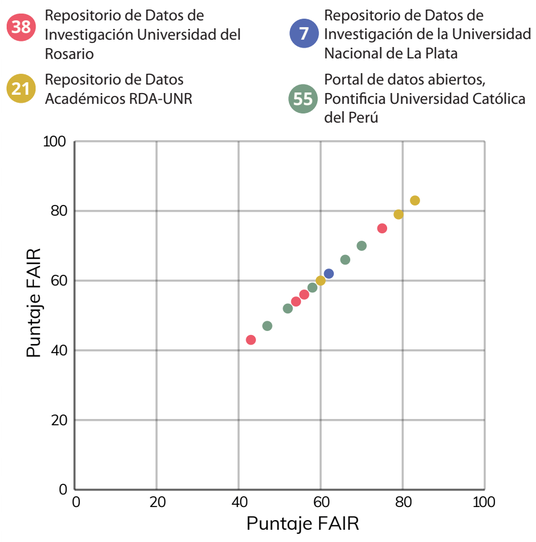
2- Distribución del puntaje FAIR en las instituciones.

3. Relación entre la clasificación dataset, puntajes FAIR y vinculación con el repositorio de datos.

Dispersión de los puntajes FAIR y coherencia de la clasificación Dataset



Repositorio de Datos de investigación
Documentos revisados



- **Evaluación FAIR**: se utilizó el evaluador FAIR www.f-uji.net/index.php?action=test

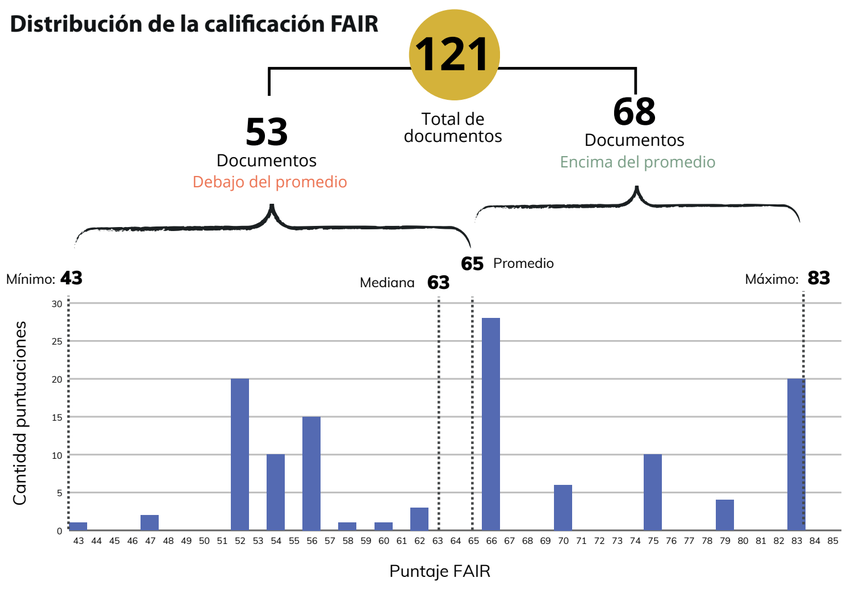
La distribución de los puntajes FAIR por documento explorado permite observar que casi un 50% de los documentos clasificados como data sets tienen puntajes inferiores a 65.

El puntaje con mayor representación es 66 puntos FAIR.

De los 121 ítems analizados, 20 tienen un puntaje superior a los 80 puntos FAIR.

Se observa que 40 documentos tienen puntuaciones ≥70 puntos FAIR, de los 121 datasets explorados.

Distribución de la calificación FAIR



Hallazgos

- Un 86% no tiene diccionario de datos, ni un readme que permita comprender y utilizar correctamente el set de datos.
- Es posible sacar una buena calificación de metadatos FAIR sin que el contenido del ítem sea un dataset.
- El 50% de los ítems analizados, tienen una calificación FAIR inferior a 66%. La característica FAIR más robusta es Encontrable. Las demás aún se encuentran incipientes.
- El 90,9% de los conjuntos de datos están abiertos, pero solo un 14% tiene una descripción o README, lo que dificulta que se puedan reutilizar con facilidad.
- Los metadatos no incluyen información sobre licencias de uso de los sets o la incluyen de forma incorrecta.
- El validador no puede identificar el formato de los archivos de los data sets para verificar si se encuentra en un formato recomendado.

Recomendaciones

- Se sugiere como requisito mínimo para los conjuntos de datos abiertos, acompañarlos de una descripción de cada una de las variables, categorías y tratamientos.
- Se sugiere reclasificar todos los ítems que se encuentran en la categoría Dataset y que no cumplen con la definición de Dataset.
- Se sugiere discutir qué formatos se pueden catalogar como universales y recomendar su uso entre los investigadores.
- Se sugiere analizar el sitio web de Kaggle (Kaggle: Your Machine Learning and Data Science Community), que es una red social, donde se promueve el uso de datos abiertos, con el fin de comprender las tendencias mundiales en formatos, accesibilidad, amigabilidad en datos.

Bibliografía

FAIR Principles. (2016). https://www.go-fair.org/fair-principles/

Hernández-Pérez, T., & García-Moreno, M. A. (2013). Datos abiertos y repositorios de datos: nuevo reto para los bibliotecarios. Profesional de la Información, 22(3). https://doi.org/10.3145/epi.2013.may.10

Morales Vargas, A., & Codina, L. (2019). Atributos de calidad web para repositorios de datos de investigación en universidades. Hipertext.net, 19. https://doi.org/10.31009/hipertext.net.2019.i19.04