



# EVALUATION OF FAIR PRINCIPLES IN RESEARCH DATA REPOSITORIES AT COLOMBIAN UNIVERSITIES

## Evaluación de principios FAIR en repositorios de datos de investigación en instituciones de educación superior en Colombia

Gineth-Andrea Lopez-Hoyos   
Universidad del Cauca (Popayán-Colombia).   
[lopezhoyos@unicauca.edu.co](mailto:lopezhoyos@unicauca.edu.co)

Sandra-Milena Roa-Martínez   
Universidad del Cauca (Popayán-Colombia).   
[smroa@unicauca.edu.co](mailto:smroa@unicauca.edu.co)

Fecha de recibido: 18-08-2024

Fecha de aceptado: 19-11-2024



### ABSTRACT

The volume of data generated in research processes represents an opportunity for collaborative work, scientific advancement, and integration into the open science movement. This is facilitated by the reuse of data for purposes such as reproducibility, among others. The main goal is to identify the landscape of research data repositories as a strategy for sharing or publishing research data from higher education institutions (HEIs) in Colombia. The research data repositories (RDR) in Colombian universities listed in the SCImago Journal Rank and the Papyrus meta-repository were analyzed, along with their respective datasets. Subsequently, the implementation of the FAIR principles (Findable, Accessible, Interoperable, Reusable) in these repositories was evaluated. It was found that although there are institutional repositories, there are few RDR. However, the initiatives to share datasets are evident when evaluating the FAIR principles, as indicated by the level of completeness found in these HEIs. It is concluded that initiatives like Papyrus in Colombia serve as platforms for indexing and making visible RDRs in Colombia, with significant contributions from HEIs. The implementation of the FAIR principles promotes the integration and sharing of data for open science.

**Keywords:** access to information; dataset; open science.

### RESUMEN

El volumen de datos generados en los procesos de investigación se constituye como una posibilidad de trabajo colaborativo, avance de la ciencia e integración al movimiento de ciencia abierta, lo cual se viabiliza a partir del reuso con fines de reproducibilidad que se puede dar a estos datos, entre otros. El objetivo de este trabajo fue identificar el panorama de los repositorios de datos de investigación (RDI) como estrategia de compartición o disponibilización de datos de investigación de las instituciones de educación superior (IES) en Colombia. Para ello, se analizó la existencia de RDI en universidades colombianas encontradas en el SCImago Journal Rank y en el meta-repositorio Papyrus y sus conjuntos de datos. Posteriormente, se evaluó la implementación de los principios FAIR (Findable, Accesible, Interoperable, Reusable) en dichos repositorios. Se encontró que, aunque existen repositorios institucionales, son pocos los RDI, no obstante, son claras las iniciativas de compartición de conjuntos de datos que se evidencian al evaluar los principios FAIR, debido al nivel de completitud encontrado en dichas IES. Se concluye que iniciativas como Papyrus en Colombia se presentan como plataformas

de indexación y visibilidad de los RDI en Colombia, con alta contribución de las IES y la implementación de los principios FAIR favorecen la integración y compartición de datos para la ciencia abierta.

**Palabras clave:** acceso a la información; ciencia abierta; conjunto de datos.

---

## 1. INTRODUCTION

Universities in Colombia generate a significant amount of high-value intellectual output for the scientific community (books, journals, articles, theses, data sets, among others), which may not be found or reused by other members of society. Duplicated efforts in similar areas could not consider the data generated by other researchers at the conceptual or practical level. Hence, there is a need for universities, particularly those that generate scientific knowledge, to share and publish this data. Additionally, these practices contribute to the research community and allow for the reproducibility and reuse of data, which are the pillars of open science[1].

Access to information produced in research is one of the main challenges researchers must face, highlighting the need for the implementation of digital ecosystems in universities across the country with the goal of fostering collaboration among researchers and the reuse of data, leading to structured advancements in the construction of science. Based on this, the objective of this work was to identify the national landscape regarding the sharing or publication of research data in the face of the growing volume of data and the open science movement, through a review of research data repositories in higher education institutions in Colombia. Additionally, it sought to determine whether the data repositories found allow their datasets to be findable, accessible, interoperable, and reusable, i.e., the incorporation of the FAIR principles (Findable, Accessible, Interoperable, Reusable), aiming to enhance and encourage practices of reuse, collaboration, and scientific advancement.

Effective metadata usage is essential to help organize and specify the information generated through research in universities, which remains unutilized due to the lack of access [2]. Metadata primarily enables the implementation of the FAIR principles.

As for research data, these are all raw data on which any research is based, and they may or may not be published when scientific progress is communicated, but they are what underpin new knowledge [3]. When this data is properly described and assigned metadata for subsequent retrieval, it is deposited in data repositories [4]. In contrast, open data is public information made available in formats that allow its use and reuse under an open license and without legal restrictions for its exploitation [1], with the aim of generating value in processes of active transparency, accountability, citizen participation, research, and innovation. Thus, data repositories are associated with the idea of valuing data as an asset of the organization, while evaluating the ability to manage, and establish responsibilities in decision-making and related tasks, ensuring data quality and proper use [5].

Data repositories could be defined as a system and set of services designed as an archive for digital data with context, stability, and persistence[6], or as a database infrastructure that collects, manages, and provides access to data, metadata, and associated documentation[4]. The material produced during the research is used to extract and validate results, promoting the idea that raw data can have a “second life” and be used beyond their initial purpose. Such uses can range from reusing data to produce new studies to serving as a means of verifying research results [7].

On the other hand, it was found that through Resolution 460 of 2022 [8], the Ministry of Information and Communication Technologies (MinTIC Colombia) issued the National Data Infrastructure Plan (PNID) and its roadmap, to drive the digital transformation of the State and the development of a data-based economy.

Some meta-repositories collect information from associated entities, such as R3Data, Datos.gov, and Papyrus, and are promoted by the national government and align with the national data infrastructure plan. R3Data is a global registry of research data repositories across various academic disciplines, promoting a culture of exchange, greater access, and better visibility of research data; it currently hosts 3,244 repositories.

Datos.gov is Colombia's national open data platform, which promotes transparency and decision-making based on public data. Meanwhile, Papyrus is a multidisciplinary repository housing scientific datasets from research projects of institutions within the Consortia Consortium, with the aim of meeting international standards and managing persistent identifiers such as the Digital Object Identifier (DOI).

According to the FAIR principles guide for scientific data management and administration, the following definitions are provided [9]: 1) **Findable**: Data and metadata should be discoverable by the community after publication, through search tools. 2) **Accessible**: Data and metadata should be accessible, allowing other researchers to download them using their identifiers. 3) **Interoperable**: Both data and metadata should be described according to community rules, using open standards, to enable their exchange and reuse. 4) **Reusable**: Data and metadata should be reusable by other researchers, with clear provenance and reuse conditions.

## 2. RELATED WORK

In Osorio et al. [1] an analysis of open data in Colombian higher education institutions (HEIs) is described, revealing that the publication of data in open format by HEIs is low, with only 5.4% of the sampled HEIs having published at least one dataset on the national open data portal. This highlights the need to encourage openness and the use of data in HEIs, establishing a framework for data governance that facilitates inter-institutional cooperation leading to knowledge generation and innovation.

Although the government has made efforts to develop open data dissemination strategies, this study confirms the low adoption of these strategies in public and private entities. It also found that further research is needed to design tools (evaluation, diagnosis, and gap identification) to support the implementation of open data initiatives and boost interinstitutional collaboration for knowledge generation, products, and services for society.

In Méndez et al. [10], a review of the status of research data repository implementation in Spanish public universities was conducted. The first step involved selecting the universities to be analyzed for the study sample. The results show the effort being made to publish datasets (1955), with 10 universities accounting for 68.3% of the published data, and an increased interest in dataset publication since 2021. This study also documents the use of technologies in the implementation of research data repositories, identifying that Dspace and Dataverse are widely used, with other platforms like EPrints, InvenioRDM, and Zenodo implemented in some of the universities studied. Using the FAIR principles, an evaluation of the aforementioned platforms was conducted, showing that Dataverse satisfies the most FAIR principles (9 fully and 4 partially), while Dspace (6 fully and 5 partially) is the most widely installed. The study concludes that the implementation of research data repositories is in an early stage.

### 3. METHODOLOGY

To analyze the research data repositories of Colombian universities, the four steps defined in [11] as follows were implemented:

**a. Definition of analysis criteria:** As analysis criteria, it was defined that the data be published in the data repository or the institutional repository. Shared data sets will be reviewed and the FAIR criteria will be evaluated. As an additional step, R3data was used to validate if the repository is indexed in this search engine.

**b. Selection of the sample population:** The SCImago Journal Rank (SJR) is a public access portal that includes journals and scientific indicators from developed countries based on information contained in the Scopus® database (Elsevier B.V.). The indicators provided can be used to evaluate and analyze scientific domains. The SJR is a classification of academic and research-related institutions based on a composite indicator that combines three different sets based on research performance, innovation results, and social impact through web visibility [12]. For this reason, the top 10 Colombian universities from the general SCImago ranking were selected to validate the existence of the data repository and assess the FAIR principles for each. The type of institution (private or public) and the thematic areas in which datasets are published, whether they have a data repository, and technologies were also identified. [Table 1](#) shows the universities selected for the validation of this study.

**Table 1.** Selected universities.

| Rank<br>(Global) | Institution                        | Type    | Institutional Repository | Data Repository |
|------------------|------------------------------------|---------|--------------------------|-----------------|
| 1(1,000)         | National University of Colombia    | Public  | X (DSpace)               | N/A             |
| 2(1,922)         | University of Antioquia            | Public  | X (DSpace)               | N/A             |
| 3(2,467)         | Pontifical Javeriana University    | Private | X (DSpace)               | N/A             |
| 4(2,472)         | University of the Andes            | Private | X (DSpace)               | Dataverse       |
| 5(3,667)         | Universidad del Rosario            | Private | X (DSpace)               | Dataverse       |
| 6(3,820)         | University of the Coast            | Public  | X (DSpace)               | Dataverse       |
| 7(4,776)         | University of Valle                | Public  | X (DSpace)               | N/A             |
| 8(4,834)         | Pontifical Bolivarian University   | Private | X (DSpace)               | N/A             |
| 9(4,903)         | University of La Sabana            | Private | X (DSpace)               | N/A             |
| 10(5,120)        | Industrial University of Santander | Public  | X (DSpace)               | N/A             |

**c. Data collection:** For this stage, the official website of each of the universities was accessed and the institutional repository was searched first. Within this repository, it was validated if there is an attached data repository or if the data sets are accessible from it. Next, each of the links available on the main page to access the institutional repository was validated. When the link was not found on the page, the link related to the institutional library was searched to access the repository from there.

**d. Analysis of data repositories:** The following activities were carried out for this analysis:

1. Accessing the university's website. Searching for the research data repository.
2. Verifying the technology implemented in the research data repository.

3. Searching within the repository for published datasets.
4. Evaluating FAIR principles for repositories that have published datasets.

In [Table 2](#), the evaluation criteria used to assess the FAIR principles are described, which help determine the current state of data repository implementation at the Colombian universities in the sample. During this analysis, it was found that few of the selected universities had datasets in a data repository, which led to the inclusion of an additional group of institutions that were identified through the snowball effect, a technique recognized in scientific research, allowing for an expanded sample to validate the FAIR principles. Therefore, it was necessary to carry out stages 3 and 4 of this methodology for this new group of universities.

**Table 2.** Evaluation Criteria. Adapted from Langer et al. [13]

| SubGroup FAIR | Criterion | Description                                                                                                                                                                    |
|---------------|-----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| FINDABLE      | F1        | Is a particular research data set in a current?                                                                                                                                |
|               | F2        | Is the research data information in that platform indexed in data catalogs, registries, and search engines?                                                                    |
|               | F3        | Is a search interface available with filter possibilities for structured Linked Data?                                                                                          |
| ACCESSIBLE    | A1        | Can new research data be stored or referenced easily?                                                                                                                          |
|               | A2        | Is the user input interface Linked Data-aware and easy to use by hiding technical terms and identifiers?                                                                       |
|               | A3        | Can the research data or metadata be accessed directly via http(s)?                                                                                                            |
|               | A4        | Do the authentication and authorization settings for public/private/restricted access exist?                                                                                   |
| INTEROPERABLE | I1        | Is the metadata description available in an RDF serialization?                                                                                                                 |
|               | I2        | Can particular established ontologies be used to describe the research dataset in a general way such as <i>schema.org/Dataset</i> , <i>Datacite</i> , <i>DCAT/DublinCore</i> ? |
|               | I3        | Can domain-specific vocabularies be used to further describe the research data set?                                                                                            |
|               | I4        | Can each concept related to the research data set be described with a corresponding URI?                                                                                       |
| REUSABLE      | R1        | Can a data license be specified in a Linked Data fashion?                                                                                                                      |
|               | R2        | Can the data provenance be specified and updated in a structured way?                                                                                                          |
|               | R3        | Are data sets in a relationship based on Linked Data and criteria such as the topic, community, used method, or similar?                                                       |
|               | R4        | Is the provided data validated or do compliance checks exist?                                                                                                                  |

The scoring system used to evaluate the FAIR principles based on the criteria is presented in [Table 3](#), which was adapted from [13]. Additionally, a value column was added to perform a comparative analysis between the data repositories of the universities in the sample.

**Table 3.** Scoring System. Adapted from Langer et al. [13]

| Indicator | Description                                                                      | Value |
|-----------|----------------------------------------------------------------------------------|-------|
| +         | If the criterion was entirely fulfilled.                                         | 1     |
| o         | If the criterion was partially fulfilled.                                        | 0.5   |
| -         | If the criterion was not fulfilled.                                              | 0     |
| ()        | If the feature is limited in the native version but might be there with plugins. | ND    |
| ?         | If it was not possible to assess the mentioned criterion.                        | NA    |
| %         | If the criterion was not applicable.                                             | NE    |

## 4. RESULTS

Once on the official page of each university from Table 1, the institutional repository and the data repository were searched, to conduct an analysis of the FAIR principles that can be adopted according to the tool used, as was done in the work by [10]. When reviewing the institutional and data repository of each university, the following were identified: the number of degree thesis, articles, datasets, and institutional documents indexed. Table 4 shows the count of the data found in the university repositories analyzed. It is worth noting that the universities were added as indicated in the methodology section and highlighted IES without datasets published.

**Table 4.** Products found in the repositories. (PY)

| Rank | Institution (Website)                                                                                                                        | TS     | PR     | DS  | ID    | R3 | DG | PY  |
|------|----------------------------------------------------------------------------------------------------------------------------------------------|--------|--------|-----|-------|----|----|-----|
| 1    | National University of Colombia ( <a href="https://repositorio.unal.edu.co">https://repositorio.unal.edu.co</a> )                            | 25,202 | 36,236 | 0   | 78    | 1  | 6  | 0   |
| 2    | University of Antioquia ( <a href="https://bibliotecadigital.udea.edu.co">https://bibliotecadigital.udea.edu.co</a> )                        | 18,244 | 701    | 0   | 1,176 | 0  | 4  | 0   |
| 3    | Pontifical Javeriana University ( <a href="https://repositorio.javeriana.edu.co">https://repositorio.javeriana.edu.co</a> )                  | 7,527  | 11,994 | 0   | 157   | 0  | 0  | 0   |
| 4    | University of the Andes ( <a href="https://repositorio.uniandes.edu.co">https://repositorio.uniandes.edu.co</a> )                            | 32,350 | 1,181  | 66  | 104   | 0  | 0  | 66  |
| 5    | Universidad del Rosario ( <a href="https://repositorio.urosario.edu.co">https://repositorio.urosario.edu.co</a> )                            | 1,273  | 11,621 | 63  | 133   | 1  | 0  | 63  |
| 6    | University of the Coast ( <a href="https://repositorio.cuc.edu.co">https://repositorio.cuc.edu.co</a> )                                      | 2,814  | 6,670  | 8   | 968   | 0  | 0  | 8   |
| 7    | University of Valle ( <a href="https://bibliotecadigital.univalle.edu.co/home">https://bibliotecadigital.univalle.edu.co/home</a> )          | 11,754 | 7,306  | 0   | 0     | 0  | 5  | 0   |
| 8    | Pontifical Bolivarian University ( <a href="https://repositorio.upb.edu.co">https://repositorio.upb.edu.co</a> )                             | 1,513  | 2,527  | 0   | 16    | 0  | 0  | 0   |
| 9    | University of La Sabana ( <a href="https://intellectum.unisabana.edu.co">https://intellectum.unisabana.edu.co</a> )                          | 8,064  | 2,769  | 0   | 2,098 | 0  | 0  | 0   |
| 10   | Industrial University of Santander ( <a href="https://noesis.uis.edu.co">https://noesis.uis.edu.co</a> )                                     | 4,059  | 5,382  | 0   | 0     | 0  | 5  | 0   |
| 11   | University Foundation of the Andean Region ( <a href="https://digitk.areandina.edu.co/home">https://digitk.areandina.edu.co/home</a> )       | 2,425  | 343    | 2   | 54    | 0  | 0  | 2   |
| 12   | Luis Amigó Catholic University ( <a href="http://repositorio.ucatolicaluis-amigo.edu.co">http://repositorio.ucatolicaluis-amigo.edu.co</a> ) | 2,550  | 2,486  | 5   | 2,005 | 0  | 0  | 8   |
| 13   | Cooperative University of Colombia ( <a href="https://repositorio.ucc.edu.co">https://repositorio.ucc.edu.co</a> )                           | 23,447 | 624    | 187 | 248   | 0  | 0  | 140 |
| 14   | University of Medellín ( <a href="https://repositorio.udem.edu.co">https://repositorio.udem.edu.co</a> )                                     | 4,404  | 1,385  | 18  | 489   | 0  | 0  | 10  |
| 15   | EAN University ( <a href="https://repositorio.universidadean.edu.co">https://repositorio.universidadean.edu.co</a> )                         | 4,221  | 9      | 1   | 1,036 | 0  | 0  | 1   |
| 16   | EIA University ( <a href="https://repositorio.eia.edu.co">https://repositorio.eia.edu.co</a> )                                               | 3,347  | 24     | 2   | 628   | 0  | 0  | 1   |
| 17   | Santo Tomás University ( <a href="https://repositorio.usta.edu.co">https://repositorio.usta.edu.co</a> )                                     | 25,892 | 7,434  | 17  | 57    | 0  | 0  | 10  |
| 18   | Externado University ( <a href="https://bdigital.uexternado.edu.co">https://bdigital.uexternado.edu.co</a> )                                 | 5,476  | 13,875 | 2   | 40    | 0  | 0  | 2   |
| 19   | University of Caldas ( <a href="https://repositorio.ucaldas.edu.co">https://repositorio.ucaldas.edu.co</a> )                                 | 1,186  | 6,468  | 0   | 33    | 0  | 2  | 0   |
| 20   | El Bosque University ( <a href="https://repositorio.unbosque.edu.co">https://repositorio.unbosque.edu.co</a> )                               | 7,691  | 2,280  | 0   | 28    | 0  | 0  | 0   |
| 21   | Higher College of Antioquia ( <a href="https://colmayor.janium.net">https://colmayor.janium.net</a> )                                        | 617    | 177    | 0   | 70    | 0  | 0  | 0   |

Thesis (TS), Paper (PR), DataSet (DS), Institutional Documents (ID), R3DATA(R3), DATA.GOV(DG), Papyrus

Subsequently, it was reviewed whether the institutional and data repositories are indexed in open data repositories such as R3data, Datos.gov, and Papyrus. The aim was to determine how many data sets were shared in these repositories that follow the national government's initiatives in relation to the publication of open data. Table 4 shows the number of datasets that the universities in the sample have made available on these platforms. After completing this review, it was found that currently, there are few universities sharing datasets and that they are simultaneously building their own data repository,

which can be accessed by the rest of the research community and contributes to the development of science.

Since only a few HEIs of the initial sample had research data repositories, which we removed from the list, and the same for the ones that did not have any data set, it was necessary to include other universities in this study to have a larger group that would allow for the evaluation of the FAIR principles in the data repositories of higher education institutions in Colombia. To achieve this, using the snowball effect, universities from the Papyrus repository were included. Table V presents the universities with a repository indexed in Papyrus that were added. It is worth noting that for the evaluation of the FAIR principles, only universities with at least one published dataset were considered. To evaluate the FAIR principles, the results matrix in Table 5 was created, based on each of the criteria that were assessed.

**Table 5.** Evaluation of FAIR criteria in repositories

| Institution                                | DS  | F |   |     | A |   |   |     | I |   |    |   | R |   |   |    |
|--------------------------------------------|-----|---|---|-----|---|---|---|-----|---|---|----|---|---|---|---|----|
|                                            |     | 1 | 2 | 3   | 1 | 2 | 3 | 4   | 1 | 2 | 3  | 4 | 1 | 2 | 3 | 4  |
| University of the Andes                    | 66  | 1 | 1 | 0.5 | 1 | 1 | 1 | 0.5 | 1 | 1 | NA | 1 | 1 | 1 | 1 | NA |
| University of the Coast                    | 8   | 1 | 1 | 0.5 | 1 | 1 | 1 | 1   | 1 | 1 | NA | 1 | 1 | 1 | 1 | NA |
| Universidad del Rosario                    | 63  | 1 | 1 | 0.5 | 1 | 1 | 1 | 0.5 | 1 | 1 | NA | 1 | 1 | 1 | 1 | NA |
| University Foundation of the Andean Region | 2   | 1 | 1 | 0.5 | 1 | 1 | 1 | 1   | 1 | 1 | NA | 1 | 1 | 1 | 1 | NA |
| Luis Amigó Catholic University             | 8   | 1 | 1 | 0.5 | 1 | 1 | 1 | 0.5 | 1 | 1 | NA | 1 | 1 | 1 | 1 | NA |
| Cooperative University of Colombia         | 140 | 1 | 1 | 0.5 | 1 | 1 | 1 | 1   | 1 | 1 | NA | 1 | 1 | 1 | 1 | NA |
| University of Medellín                     | 10  | 1 | 1 | 0.5 | 1 | 1 | 1 | 1   | 1 | 1 | NA | 1 | 1 | 1 | 1 | NA |
| EAN University                             | 1   | 1 | 1 | 0.5 | 1 | 1 | 1 | 1   | 1 | 1 | NA | 1 | 1 | 1 | 1 | NA |
| EIA University                             | 2   | 1 | 1 | 0.5 | 1 | 1 | 1 | 0.5 | 1 | 1 | NA | 1 | 1 | 1 | 1 | NA |
| Santo Tomás University                     | 10  | 1 | 1 | 0.5 | 1 | 1 | 1 | 0.5 | 1 | 1 | NA | 1 | 1 | 1 | 1 | NA |
| Externado University                       | 2   | 1 | 1 | 0.5 | 1 | 1 | 1 | 0.5 | 1 | 1 | NA | 1 | 1 | 1 | 1 | NA |

Each of the FAIR principles was determined for the universities' repositories. The results indicate whether the responses to each of the questions regarding the datasets in these repositories are: Findable (F), Accessible (A), Interoperable (I), and Reusable (R), following the scoring system described in the methodology.

A total of 11 universities were evaluated, including the 4 universities selected from the SCImago ranking, and the universities added through the snowball sampling. Those that did not have datasets available in Papyrus, such as Javeriana University, El Bosque University, University of Caldas, and Higher College of Antioquia were excluded at the time the data was collected.

## 5. DISCUSSION

Based on the results, it can be observed that the implementation of data repositories in Colombian universities is still in the initial phase. All the analyzed universities have an institutional repository that provides access to relevant information for researchers. In contrast, data repositories have only been implemented by 3 out of the 11 universities initially evaluated, demonstrating that despite national government directives, universities are not aligned with open data and the sharing of research data suggested by the Open Science movement. Additionally, on some of these institutions' websites, it was difficult to find the link associated with the institutional repository. In most cases (80%), one must access the library or use a search engine to directly access the institutional or data repository.

Figure 1 shows that only two (2) universities, Universidad del Rosario and the National University of Colombia, have their data repositories indexed in R3data, offering the published information to the entire research community, thereby achieving greater visibility. At the same time, it is observed that only 4 universities from the sample have published a dataset on the DATOS.GOV website, which mainly corresponds to selection processes and mandatory financial reports for public entities. It is worth highlighting that none of the analyzed private universities publish datasets on this portal.

Concerning the technology implemented for institutional repositories, it was found that the universities in the initial sample, selected based on the SCImago ranking, use DSpace, primarily due to its ease of implementation[13]. Therefore, it is important to highlight the FAIR principles that this technology can support and its main limitations. According to [13], DSpace fully complies with 7 and partially with 4 of the 15 evaluated criteria. The four remaining criteria that are not met correspond to I1, I3, and I4, which are part of the INTEROPERABLE criterion and refer to the availability of metadata descriptions in RDF, the use of specific vocabularies, and the description of each concept related to the dataset, respectively.

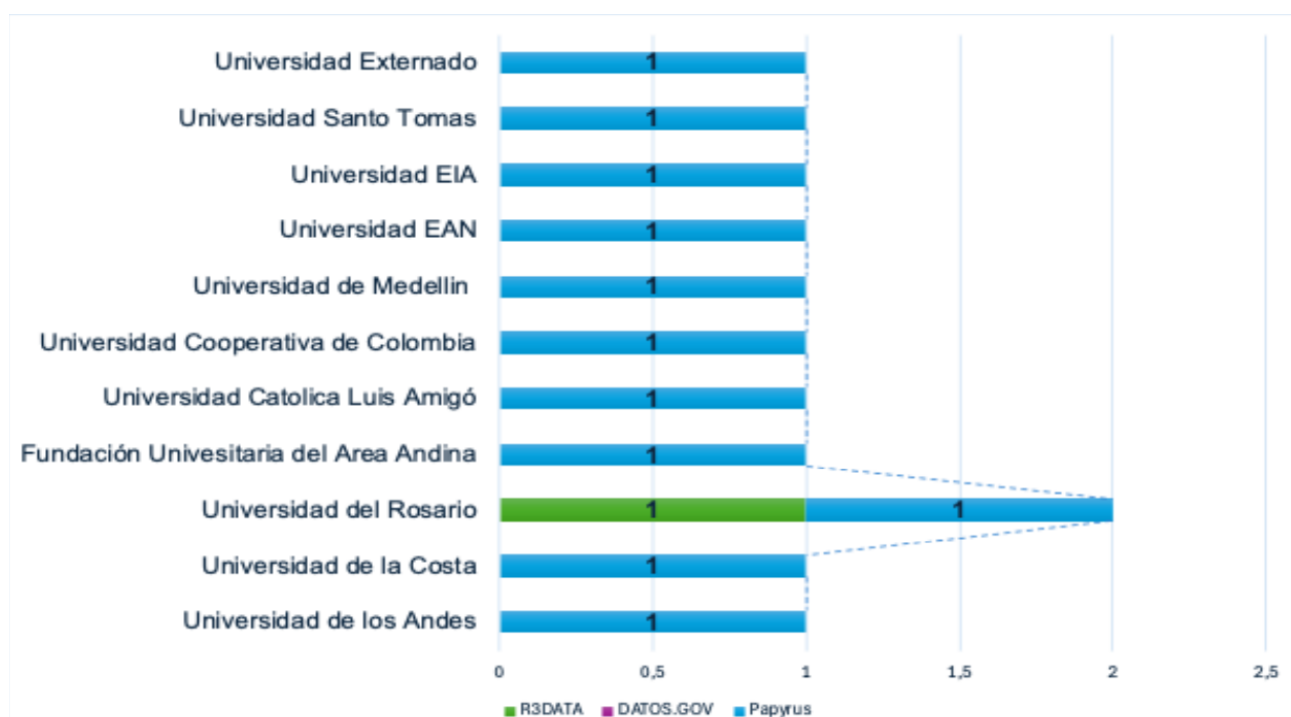


Figure 1. Open data initiatives.

Lastly, R4 evaluates whether the provided data are validated as part of the REUSABLE criterion. Dataverse fully complies with 9 of the 15 evaluated criteria and partially with 4. Only two criteria are not met, I3 y R4, which coincide with two of the criteria that DSpace also fails to meet, corresponding to the use of specific vocabularies and data validation. Dataverse is the technology implemented in the Papyrus metadata repository from which they took the snowball sample.

Regarding the review of the shared data in the repositories —theses, articles, datasets, and institutional documents—, it was found that the largest volume of data corresponds to theses (60%) and research articles (37%), accounting for 97% of the total data found. On the other hand, datasets make up only 0.11%. The Cooperative University of Colombia stands out with 140 datasets, the Universidad del Rosario, with 67 datasets, and the University of the Andes, with 63 datasets in its data repository.

Concerning the thematic areas of the datasets, 32% correspond to Social Sciences, 20% to Medicine, Health, and Life Sciences, and 14% to Business and Administration. It was also found that areas such as Engineering and Computing and Information Sciences only contribute 9% and 7%, respectively, even though many processes require data collection and processing, which would suggest a higher number of datasets in these areas. Particularly, it is noted that, for example, the Cooperative University of Colombia contributes a high number of datasets in Medicine, Health, and Life Sciences, highlighting the publication of 37 datasets in 2023, compared to 2 datasets published in the current year, indicating a decrease in the dataset publication process.

Regarding the FAIR principles, [Figure 2](#) shows that since all datasets are available in Papyrus, they meet most of the criteria at the same percentage. The variations correspond to the use of more detailed metadata and the existence of access restrictions to the datasets. Concerning the FIND principle, it can be observed that all universities have a score of 2.5, corresponding to 2 criteria fully met and one partially met. This ensures that the data have a DOI, are indexed in data catalogs, and allow for data filtering, effectively supporting the principle of findability (data and metadata can be found by the community after being published through a search engine). The findability evaluated in this section indicates that the data and metadata can be accessed by the community through search tools, with the metadata including descriptive information about the context or characteristics of the data.

About the ACCESSIBLE principle, which aims to ensure that data and metadata are accessible and can be downloaded by other researchers using their identifiers, a score of 3.5 was achieved for 54% of the universities in the sample. The remaining 46% corresponds to universities that have restrictions on accessing these data within the repository, which is why these universities scored 4 points. The criterion that was fully met and scored 1 point corresponds to A4, which validates whether there are authentication parameters for data usage. Data access can be carried out via a standard, open, free, and universally implementable communication protocol, and if necessary, there will be a system for data authentication and authorization. Access to metadata will remain even when the data is no longer available.

For the INTEROPERABLE principle, a score of 3 was given to all the universities in the sample, indicating that three criteria were fully met, and one criterion, I3, could not be evaluated. This criterion corresponds to the use of domain-specific vocabularies. This does not mean that the goal of the criterion ensuring that both data and metadata are described according to community rules using open standards to allow for their exchange and reuse is not covered. It only suggests that new functionalities should be implemented to enable the use of metadata standards such as (schema.org/Dataset, DataCite, DCAT/DublinCore), which focus on the interoperability of datasets, allowing their use in various tasks related to the thematic area. This grants researchers the possibility of reusing and reproducing experiments conducted by others, thus enhancing accessibility, enabling the sharing of information obtained from studies, and fostering active collaboration in innovation, solution creation, and reinforcing research results with the same focus.

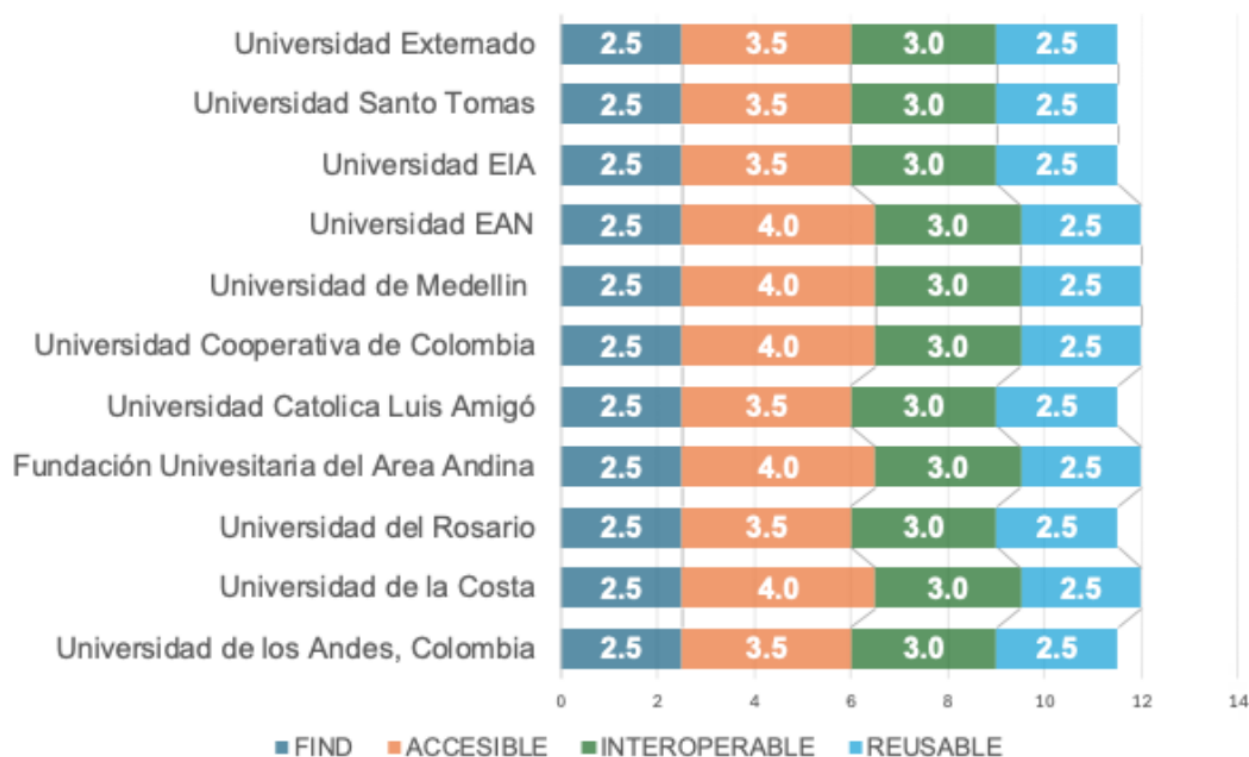


Figure 2. FAIR principles evaluation

Finally, the REUSABLE principle, which helps ensure that data and metadata can be reused by other researchers, scored 2.5 in 100% of the sample. This indicates that two criteria, R1 and R2, are fully met, while R3 is partially met, and R4 could not be evaluated as it corresponds to data compliance controls that are not implemented in the technology. The reusability of data is one of the most important focuses in the open science movement, where linked data has a significant impact due to the possibility of linking datasets through their associated metadata. This allows researchers to find other datasets that can be used, thereby adding value to new research conducted. Due to the limitations of the technology implemented for data repositories, criteria 10 and 15 were not evaluated in this review, which is why they appear as NA (Not Applicable) in the results section table.

In conclusion, it can be affirmed that the implementation of the FAIR principles has been gradually developed in the repositories of higher education institutions, laying the foundation for a more detailed implementation that supports the guidelines established to achieve the main goals of open science, namely data reuse and online accessibility.

## 6. CONCLUSIONS

From the review conducted, it was found that the universities in the sample have institutional repositories. However, few of them have datasets available in these repositories, while others have implemented research data repositories that are indexed in meta-repositories such as Papyrus. This demonstrates the progress of higher education institutions in Colombia in terms of sharing and publishing data in open science.

The results of the evaluation of the FAIR principles show that there are differences between the technologies implemented for data repositories and the definitions of the FAIR principles, particularly regarding the inclusion of Linked Data, domain-specific vocabularies, and, from another perspective, the controls for both access and validation of the provided data.

Papyrus is an initiative in which many universities have partnered to pursue the common goal of publishing research datasets. For universities that do not have a data repository, joining this initiative could become an implementation platform focused on consolidating institutional efforts in this area.

For future work, initiatives such as reviewing the FAIR criteria that are partially implemented, not implemented, or have not been evaluated are proposed, to improve the data repositories of Colombian universities, especially those related to Linked Data and specific vocabularies.

## AUTHORS' CONTRIBUTION

**Gineth Andrea Lopez-Hoyos:** Conceptualization; Methodology; Investigation; Formal Analysis; Writing-review and editing.

**Sandra Milena Roa-Martínez:** Conceptualization; Supervision; Writing-review and editing.

## ACKNOWLEDGMENTS

We sincerely express our gratitude to the University of Cauca, especially to the Artificial Intelligence Research Group (GICO) for their support in developing this work.

## REFERENCES

- [1] M.-A. Osorio-Sanabria, F. O. Amaya Fernández, M. González-Zabala, "Análisis de datos abiertos de instituciones de educación superior colombianas como apoyo a la relación Universidad-Entorno," *Entramado*, vol. 16, no. 1, pp. 272-284, Dec. 2020.  
<https://doi.org/10.18041/1900-3803/entramado.1.6127>
- [2] P. Hartmann, J. Henkel, "The rise of corporate science in AI: Data as a strategic resource," *Academy of Management Discoveries*, vol. 6, no. 3, pp. 359-381, Oct. 2020. <https://doi.org/10.5465/amd.2019.0043>
- [3] S. M. Roa-Martínez, S. A. Vidotti, R. C. Santana, "Estructura propuesta del artículo de datos como publicación científica," *Revista Española de Documentación Científica*, vol. 40, no. 1, e167, Mar. 2017.  
<https://doi.org/10.3989/redc.2017.1.1375>
- [4] S. Kowalczyk, K. Shankar, "Data sharing in the sciences," *Annual Review of Information Science and Technology*, vol. 45, pp. 247-294, 2011. <https://doi.org/10.1002/aris.2011.1440450113>
- [5] M. Salvador, C. Ramie, "Capacidades analíticas y gobernanza de datos en la Administración pública como paso previo a la introducción de la Inteligencia Artificial," *Revista del CLAD Reforma y Democracia*, no. 77, pp. 5-36, 2020.
- [6] S. M. Angelozzi, "La gestión de datos de investigación en abierto: introducción al rol emergente para las bibliotecas universitarias y científicas argentinas," *Palabra Clave (La Plata)*, vol. 9, no. 2, e091, Apr. 2020. <https://doi.org/10.24215/18539912E091>

- [7] R. Aleixandre-Benavent, R. Lucas-Domínguez, A. Sixto-Costoya, A. Vidal-Infer, "The sharing of research data in the cell & Tissue engineering area: Is it a common practice?," *Stem Cells and Development*, vol. 27, no. 11, pp. 717-722, Jun. 2018. <https://doi.org/10.1089/SCD.2018.0036>
- [8] MinTIC, *MinTIC expide el Plan Nacional de Infraestructura de Datos, que impulsará la transformación digital del Estado*, 2024. <https://goo.su/QEUWUy>
- [9] M. D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, pp. 1-9, Mar. 2016. <https://doi.org/10.1038/sdata.2016.18>
- [10] F. Méndez, A. Baptista, R. López, A. Vázquez, "Implementación de los repositorios de datos de investigación en las universidades públicas españolas: estado de la cuestión," *Scire*, vol. 29, no. 2, pp. 39-49, Nov. 2023. <https://doi.org/10.54886/scire.v29i2.4914>
- [11] M. I. S. Oliveira, H. R. de Oliveira, L. A. Oliveira, B. F. Lóscio, "Open Government Data Portals Analysis," in *Proceedings of the 17th International Digital Government Research Conference on Digital Government Research*, New York, NY, USA: ACM, 2016. <https://doi.org/10.1145/2912160.2912163>
- [12] SCImago, *SJR - SCImago Journal & Country Rank*, 2024. <https://www.scimagojr.com/aboutus.php>
- [13] A. Langer, E. Bilz, M. Gaedke, *Analysis of current RDM applications for the interdisciplinary publication of research data*, 2019. <https://ceur-ws.org/Vol-2447/paper1.pdf>