

# Análisis de cumplimiento de principios FAIR en los reportados en el sistema de hoja de vida de investigadores colombiano- CVLAC

Juan Sebastián González Sanabria<sup>1</sup>, Elena Verdú<sup>2</sup>, Xiomara Blanco<sup>2</sup>, José Texeir<sup>3</sup>, Thomas Sorza Sierra<sup>1</sup>

<sup>1</sup> Universidad Pedagógica y Tecnológica de Colombia  
Tunja, Colombia

<sup>2</sup> Universidad Internacional de La Rioja  
La Rioja, España

<sup>3</sup> Latin American and Caribbean Consortium of Engineering Institutions  
Washington, Estados Unidos

## Resumen

Los sistemas de información sobre los currículos de investigadores se han convertido en una fuente de datos para que los gobiernos puedan tomar decisiones estratégicas respecto a las inversiones. Sin embargo, estos sistemas resultan insuficientes cuando la información es incorrecta o incompleta. Este trabajo presenta una evaluación del cumplimiento del principio FAIR en los datos registrados en el sistema de currículos de investigadores colombianos (CvLAC). Para ello, se realizó un trabajo centrado en la extracción, manipulación y depuración de los datos del sistema, con el fin de realizar análisis, estudios y validaciones con sistemas internacionales como Crossref, encontrando errores notables en forma de datos inconsistentes o mal reportados, como autores y datos que no corresponden a artículos o productos existentes. Identificar estos errores contribuye a optimizar los procesos de evaluación de la investigación y, por lo tanto, a una mejor toma de decisiones estratégicas.

**Palabras clave:** CvLAC; principio FAIR; veracidad de la información

## Abstract

*Researcher resume information systems have become a source of data for governments to make strategic investment decisions. However, these systems are insufficient when the information is*

*incorrect or incomplete. This paper presents an assessment of compliance with the FAIR principle in the data recorded in the Colombian Researcher Resume System (CvLAC). To this end, work focused on the extraction, manipulation, and cleaning of the system's data. This was done to conduct analyses, studies, and validations with international systems such as Crossref. Notable errors were identified in the form of inconsistent or misreported data, such as authors and data that do not correspond to existing articles or products. Identifying these errors contributes to optimizing research evaluation processes and, therefore, to better strategic decision-making.*

**Keywords:** CvLAC; FAIR principle; accuracy of information

## 1. Introducción

A nivel global, existen diversos sistemas, tanto gubernamentales como de organizaciones, dedicados a la creación de hojas de vida para investigadores. Estos sistemas se desarrollan siguiendo características comunes, tales como (Guenther, 2003; iDecor, 2023):

1. **Estándar de metadatos:** El sistema utiliza formatos estandarizados, como Dublin Core, para almacenar la información investigativa.
2. **Validación:** La información registrada en la plataforma se somete a un proceso de análisis o evaluación.
3. **Integración con sistemas internacionales:** Facilita la conexión con plataformas como ORCID, SciELO, LILACS, Scopus, entre otras.
4. **Exportación:** Permite exportar los datos disponibles en formatos digitales.
5. **Datos registrados en un artículo:** Los productos científicos, especialmente los artículos, siguen una estructura estandarizada para su registro.

En una revisión realizada sobre los sistemas de hojas de vida de investigadores en América Latina, se observó que la mayoría no cumple con el conjunto completo de las métricas establecidas, particularmente en áreas críticas como la validación adecuada de la información, a través de la comparación o vinculación con identificadores internacionales reconocidos a nivel de artículo o autor. Asimismo, se pudo evidenciar que países como Perú y Paraguay aplican estándares de metadatos, lo que facilita la extracción y análisis de los datos reportados.

En Colombia, el sistema CvLAC incluye información sobre conferencias, artículos, proyectos y otras actividades de investigación. Tras analizar esta plataforma, se identificaron problemas relacionados con su estructura de datos, así como dificultades para verificar la calidad y veracidad de la información registrada (García-Cepero, 2010; González-Zabala et al., 2017). En el presente trabajo se analizará exclusivamente lo relacionado a la información reportada para los productos de generación de nuevo conocimiento clasificados como artículos de investigación.

## 2. Metodología

En una fase exploratoria de la plataforma, se encuentra que, para el registro de artículos, CvLAC permite al usuario diligenciar la información básica de cada artículo, en la Tabla 1 se encuentran los diferentes metadatos descriptivos que se solicitan al momento de su registro.

Tabla 1. Campos registrados para los artículos en la plataforma CVLAC.

| <b>Campo</b>               | <b>Descripción</b>   |
|----------------------------|--|
| <b>Tipo de artículo</b>    | Clasificación dada por el Ministerio a los artículos. Por ejemplo: Publicado en revista especializada, artículo corto, entre otros |
| <b>Autores</b>             | Lista de los autores del artículo, separados por coma  |
| <b>Título del artículo</b> | Título del artículo tal y como aparece en la publicación   |
| <b>País</b>                | País de publicación del artículo   |
| <b>Año</b>                 | Año de publicación   |
| <b>Página inicio</b>       | Página de inicio en la revista   |
| <b>Página fin</b>          | Página final del artículo en la revista  |
| <b>Revista</b>             | Nombre de la revista en que se publicó el artículo   |
| <b>ISSN</b>                | ISSN de la revista en la que se publicó el artículo  |
| <b>Fascículo</b>           | Fascículo de la revista  |
| <b>Volumen</b>             | Volumen de la revista  |
| <b>Editorial</b>           | Casa Editorial a la que pertenece la revista   |
| <b>DOI</b>                 | Identificador DOI del artículo   |
| <b>Palabras</b>            | Palabras clave del artículo  |
| <b>Sectores</b>            | Sectores en los que se trabaja el artículo, por ejemplo: Salud Humana-Cuidado a la salud de poblaciones humanas                    |
| <b>Reconocimiento</b>      | Reconocimientos que ha tenido el artículo  |

Para el análisis de la información, y el desarrollo de la investigación se hizo uso de una metodología exploratoria con los siguientes pasos:

- Extracción de datos. Se evaluó la estructura en la que se presentaba la información de los currículos en la plataforma CvLAC para la extracción de datos. Dada la cantidad de información registrada en los currículos, se optó por tener en cuenta: los datos personales registrados del autor y los artículos mediante el desarrolló una herramienta específica para la extracción automatizada de los datos con el uso de Python y el framework scrapy, nombrada "CvLAC Scraper".
- Limpieza y transformación de Datos. Se definió un estándar de representación de los datos en formato JSON. Luego, durante el proceso de limpieza se utilizaron expresiones regulares para eliminar inconsistencias presentes en los datos extraídos como caracteres no deseados, corrección de formatos inconsistentes y normalización de la información.
- Validación de datos. En esta fase, se realizó un análisis exploratorio de datos sobre los datos extraídos y se establecieron métricas para evaluar la consistencia y completitud de los datos en relación con los campos relevantes. Posteriormente, se llevó a cabo la verificación de los DOI de los artículos científicos mencionados en los currículos mediante el uso de la API provista. De igual modo, se verificó el identificador de autor ORCID.

Finalmente, se realizó una comparación de la información extraída de los currículos con fuentes internacionales relevantes.

### 3. Resultados

La Figura 1 ilustra los elementos de la herramienta desarrollada llamada CvLAC Scraper desde su configuración, ejecución y salida de datos para el proceso inicial de extracción de datos desde la plataforma.

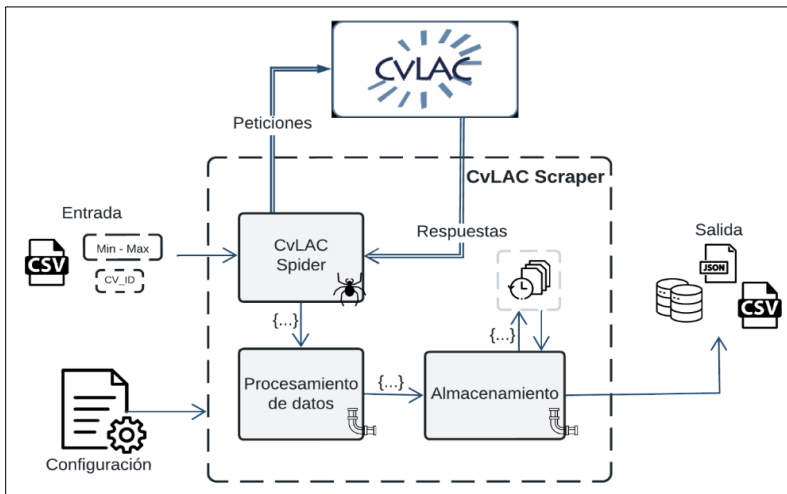


Figura 1. Arquitectura de CvLAC Scraper.

CvLAC Scraper cuenta con un archivo de configuración inicial para poder extraer la información de los sitios de los currículos. Allí se definen aspectos como: la URL de los currículos de CvLAC, el tipo de entrada de los identificadores de cada currículo, la persistencia de los datos, entre otros (Figura 2).

```

URL_BASE = 'https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh='
ID_DIGITS_LEN = 10
INPUT_TYPE = 'RANGE'
MIN_RANGE = 0
MAX_RANGE = 2117500
STORAGE_OBJECT = util.storage.MongoDBStorage()
INPUT_IDS_FILE = ''
OUTPUT_FILE = ''
LOG_FILE = 'logs/errors.txt'
URL_WITH_ERRORS_FILE = 'logs/url_errors.txt'
STORING_BATCH_SIZE = 10000
    
```

Figura 2. Archivo de configuración de CvLAC Scraper.

Dependiendo del tipo de entrada que se estableciera en el archivo de configuración, el spider recibe los identificadores a utilizar, siendo esta entrada un archivo CSV, un identificador único o un rango de identificadores. En el caso del rango, se hace una petición a todos los currículos con



id dentro del rango, incluyendo los extremos, es decir los valores menores y mayores del rango, para el caso de la extracción de los datos se usó un rango de 0 a 2117500.

Al iniciar la ejecución del Scraper, el spider de CvLAC lee el identificador o identificadores que se hayan enviado como parámetro, lanzando un número determinado de peticiones simultáneas a la página de CvLAC, conteniendo cada petición la URL completa al currículum a consultar. Luego de esto, CvLAC retorna el HTML del currículum solicitado y el spider envía esta petición al pipeline de procesamiento de datos para extraer la información relevante de dicho HTML.

El pipeline de procesamiento de datos se encarga de extraer los datos del currículum de un autor haciendo uso de las expresiones regulares, consultas XPATH y la estructura definida, una vez completada la extracción de los datos de un currículum, este pipeline envía los datos procesados al siguiente pipeline en forma de ítems o diccionarios.

Luego de procesar los datos de un currículum, el ítem con los datos estructurados pasa al pipeline de almacenamiento, en este los ítems se guardan en memoria hasta completar un número determinado para iniciar el proceso de persistencia y liberar la memoria de estos ítems y seguir almacenando los ítems que vayan llegando. Resultado de esto, se obtiene de los datos extraídos de la plataforma un total de 2.117.500 identificadores de perfiles, con tan solo 995.822 de ellos vinculados a un currículum.

Al examinar los datos extraídos de CvLAC, se observó que, de los artículos registrados, menos del 50% tiene un DOI vinculado. En cuanto a los identificadores ORCID registrados, se utilizó la API de ORCID para validar que los datos correspondieran con un ORCID existente, asegurando que el autor registrado en CvLAC coincidiera con el identificado por el sistema ORCID.

En relación con los DOI, se analizaron 790.742 artículos registrados (Figura 3), de los cuales el 37,42% tenía un DOI asociado. Sin embargo, solo el 20,60% de estos DOIs eran válidos y verificables a través de la API de la DOI Foundation, y apenas el 13,15% contenía un DOI único y válido, dado que se encontraron casos en los que varios artículos en CvLAC compartían el mismo DOI.

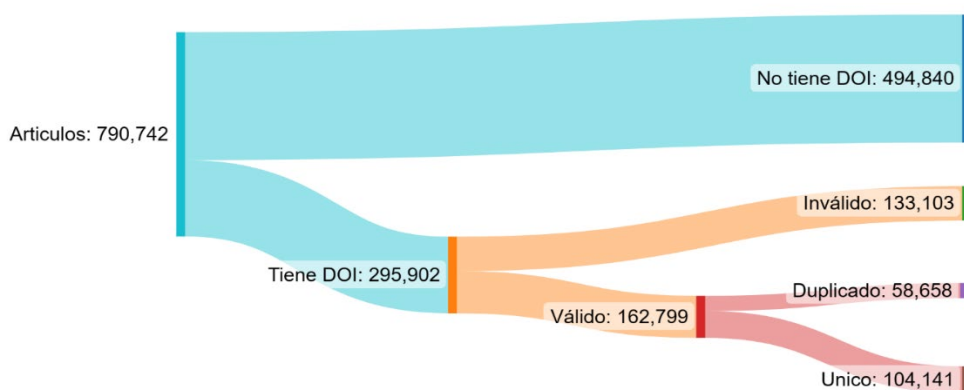


Figura 3. Comportamiento de los artículos y el DOI reportado.

Se utilizó Crossref para validar el principio FAIR de los datos, empezando con dificultades en su veracidad, dado que al validar los datos registrados en los artículos mediante el identificador DOI, accediendo a la información correspondiente en la base de datos de Crossref, se encuentran serias inconsistencias. Por ejemplo, al verificar el porcentaje de coincidencia entre los títulos de los artículos en CvLAC y los de Crossref. Los metadatos con menor porcentaje de coincidencia fueron los de **Resource Identifier**, debido a la cantidad de artículos sin DOI, lo que impidió su verificación. Además, los metadatos **Subject** y **Keywords** mostraron una baja puntuación debido a que no se obliga a registrar estos datos en CvLAC.

#### 4. Conclusiones

Esta investigación permitió concluir que, aunque CvLAC facilita la consulta y el registro de información de los investigadores y sus currículos, existen áreas de mejora en cuanto a su funcionalidad. Entre las posibles mejoras se incluye la implementación del estándar Dublin Core para estandarizar la estructura de los datos, la aplicación de criterios de validación para los identificadores de autores y recursos, y el fortalecimiento de la plataforma mediante una mayor integración con aplicaciones externas a través de un API.

Se realizó de manera satisfactoria la evaluación de cumplimiento de los principios FAIR, sintetizando los siguientes hallazgos:

- Encontrables (Findable): Los datos carecen de una difícil recuperación, al no hacer uso de estándares de metadatos se dificulta registrar o indexar los datos correctamente en buscadores.
- Accesibles (Accessible): Si bien los datos están disponibles para consulta, no pueden ser descargados por los interesados. Se sugiere utilizar formatos de comunicación abiertos, gratuitos y de dominio por la comunidad, por ejemplo, hojas de cálculo.
- Interoperables (Interoperable): Los datos no cumplen los estándares de la comunidad científica como Dublin Core, lo que hace que no se usen vocabularios y metadatos para ser reconocidos internacionalmente.
- Reutilizables (Reusable): Se sugiere publicar los datos con atributos precisos y relevantes, asociarlos con una licencia de uso y reutilización.

#### 5. Referencias

- iDeco. (2023). *Principios FAIR: cómo hacer tus datos accesibles e interoperables*. Consultado el 1 de diciembre de 2024 en <https://www.idecor.gob.ar/principios-fair-como-hacer-tus-datos-accesibles-e-interoperables/>
- García-Cepero, M. C. (2010). El estudio de la productividad académica de profesores universitarios a través de análisis factorial confirmatorio; el caso de psicología en Estados Unidos de América. *Universitas Psychologica*, Vol. 9, No. 1, pp. 13-26. <https://doi.org/10.11144/Javeriana.upsy9-1.epap>
- González-Zabala, M., Galvis-Lista, E., and Angulo-Cuentas, G. (2017). Análisis de indicadores de ciencia, tecnología e innovación (CTI) propuestos por organizaciones nacionales de CTI en América Latina. *Revista Virtual Universidad Católica del Norte*, Vol. 52, pp. 23-40.

- Guenther, R. (2003). MODS: the metadata object description schema. *Libraries and the academy*, Vol. 3, No. 1, pp. 137-150. <https://doi.org/10.1353/pla.2003.0006>

## Sobre los autores

- **Juan Sebastián González Sanabria:** Ingeniero de Sistemas y Computación, Máster en Ingeniería del Software. Estudiante Doctorado en ciencias de la computación en la Universidad Internacional de La Rioja. Profesor de la Universidad Pedagógica y Tecnológica de Colombia. [juansebastian.gonzalez@uptc.edu.co](mailto:juansebastian.gonzalez@uptc.edu.co). ORCID: <https://orcid.org/0000-0002-1024-6077>
- **Elena Verdú:** Doctora en Educación. Profesora Universidad Internacional de La Rioja, La Rioja. [elena.verdu@unir.net](mailto:elena.verdu@unir.net). ORCID: <https://orcid.org/0000-0002-3040-7077>
- **Xiomara Blanco:** Doctora en Computación. Profesora Universidad Internacional de La Rioja. [xiomarapatricia.blanco@unir.net](mailto:xiomarapatricia.blanco@unir.net)
- **José Texeir:** Doctor en Ciencias Informáticas. Director Asistente de Latin American and Caribbean Consortium of Engineering Institutions-LACCEI. [texier@laccei.org](mailto:texier@laccei.org)
- **Thomas Sorza Sierra:** Estudiante de Ingeniería de Sistemas y Computación de la Universidad Pedagógica y Tecnológica de Colombia. [thomas.sorza@uptc.edu.co](mailto:thomas.sorza@uptc.edu.co)

---

Los puntos de vista expresados en este artículo no reflejan necesariamente la opinión de la Asociación Colombiana de Facultades de Ingeniería.

Copyright © 2025 Asociación Colombiana de Facultades de Ingeniería (ACOFI)