

Calidad e integridad de los datos usados para la medición de la producción científica. Caso de estudio: CVLAC (Colombia)

Juan-Sebastián González-Sanabria; Esteban Novoa-Quinónez; Samuel-Felipe Ruiz

juansebastian.gonzalez@uptc.edu.co



Uptc[®]
Universidad Pedagógica y
Tecnológica de Colombia

Agenda

01

Introducción

02

Problemática

03

Metodología

04

Resultados

05

Conclusiones

Introducción



Busca facilitar la validación de datos de investigación en CvLAC



A través de la extracción, manipulación y limpieza de información



Permitiendo el análisis de datos sobre investigadores y artículos registrados en el sistema.

Problemática

La falta de una estructuración adecuada para el registro de información científica por parte del sistema CVLAC dificulta la validación, veracidad y estudio de la información científica registrada en el sistema dificultando el proceso de estudio de la producción científica.

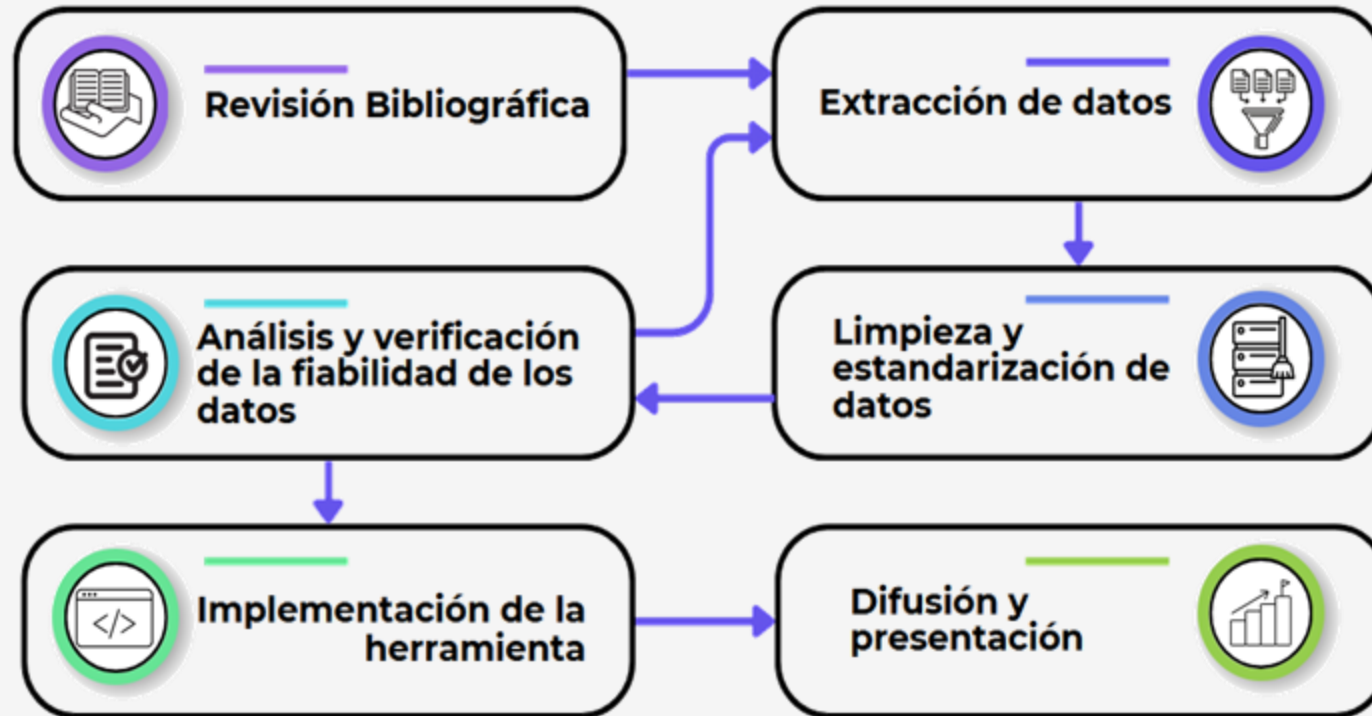


Problemática

- Implementar un mecanismo de extracción de datos de forma automática en la plataforma.
- Analizar el cumplimiento de principios FAIR en los datos reportados en el sistema CvLAC.
- Proponer alternativas de mejora de la publicación de datos CvLAC.



Metodología



Revisión Contextual

Se realizaron consultas a diversas fuentes bibliográficas para recopilar información de los trabajos previos sobre la validación de la producción científica.



Revisión Contextual



FINDABLE



ACCESSIBLE



INTER-
OPERABLE

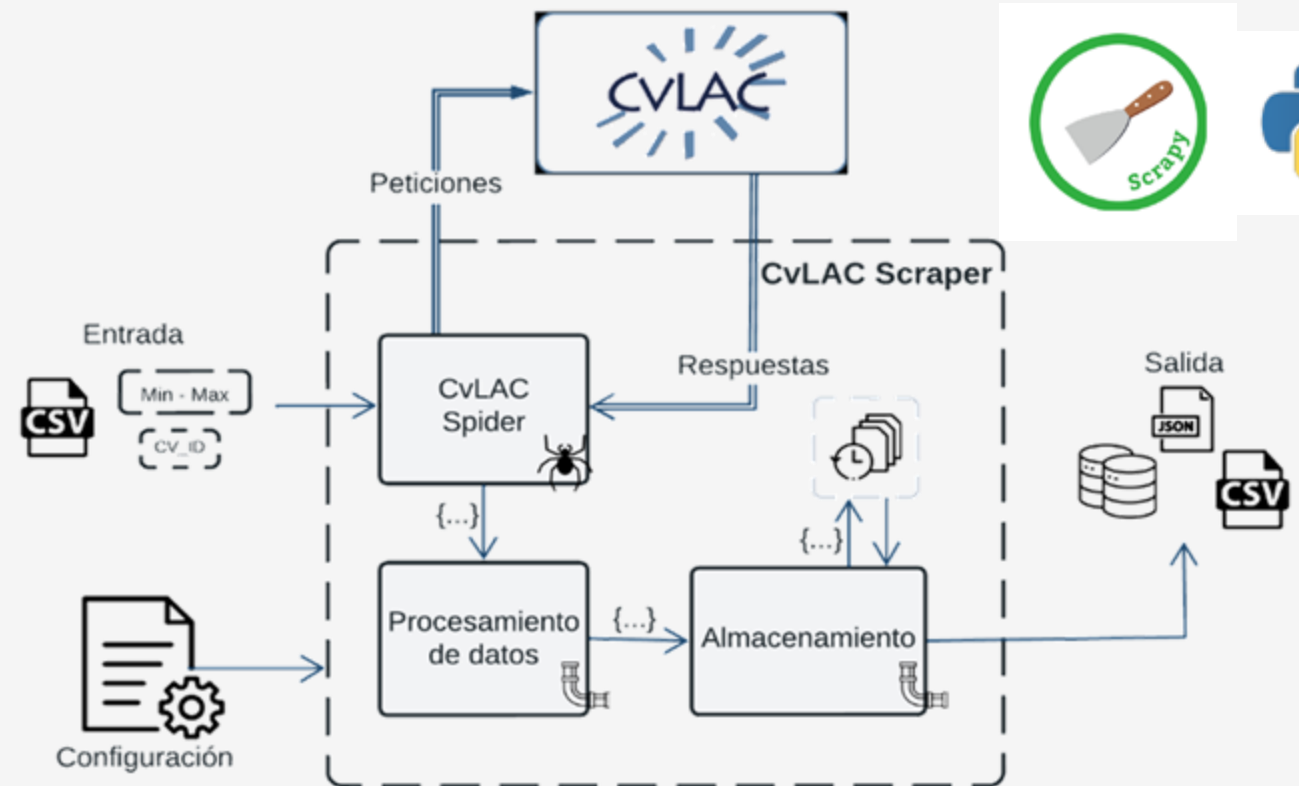


REUSABLE



Extracción de los datos

- Evaluación estructura de los currículos de CvLAC
- Definición de consultas XPath relevantes para la extracción
- Automatización y consulta masiva de currículos
- Desarrollo de la herramienta de extracción



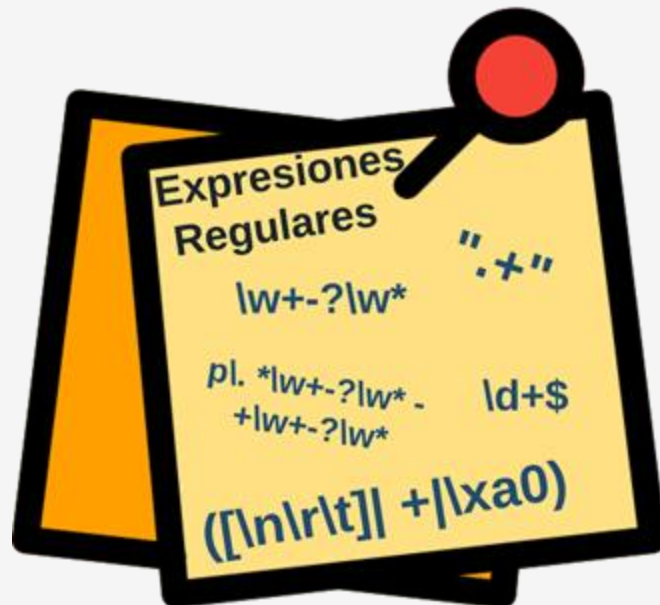
URL de un currículo:

https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0001344275

Limpieza y estandarización de los datos

Uso de expresiones regulares para:

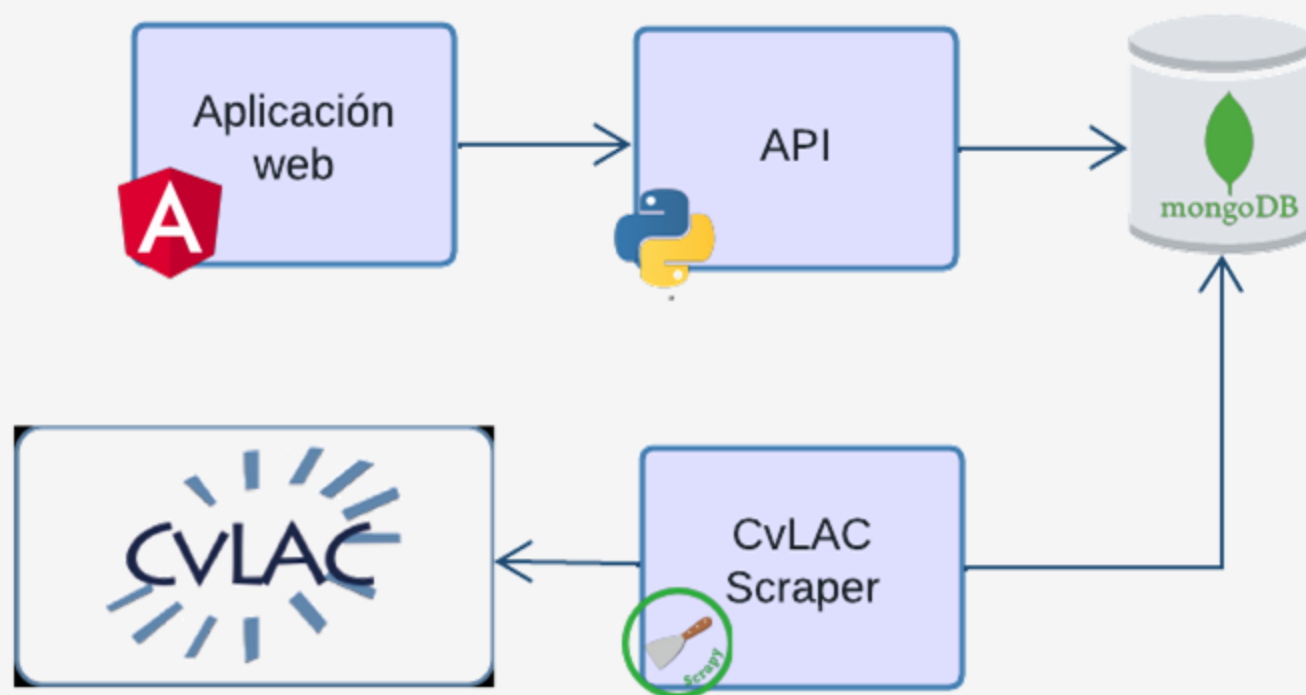
- Extraer información de currículos
- Limpieza de caracteres
- Estandarización de campos



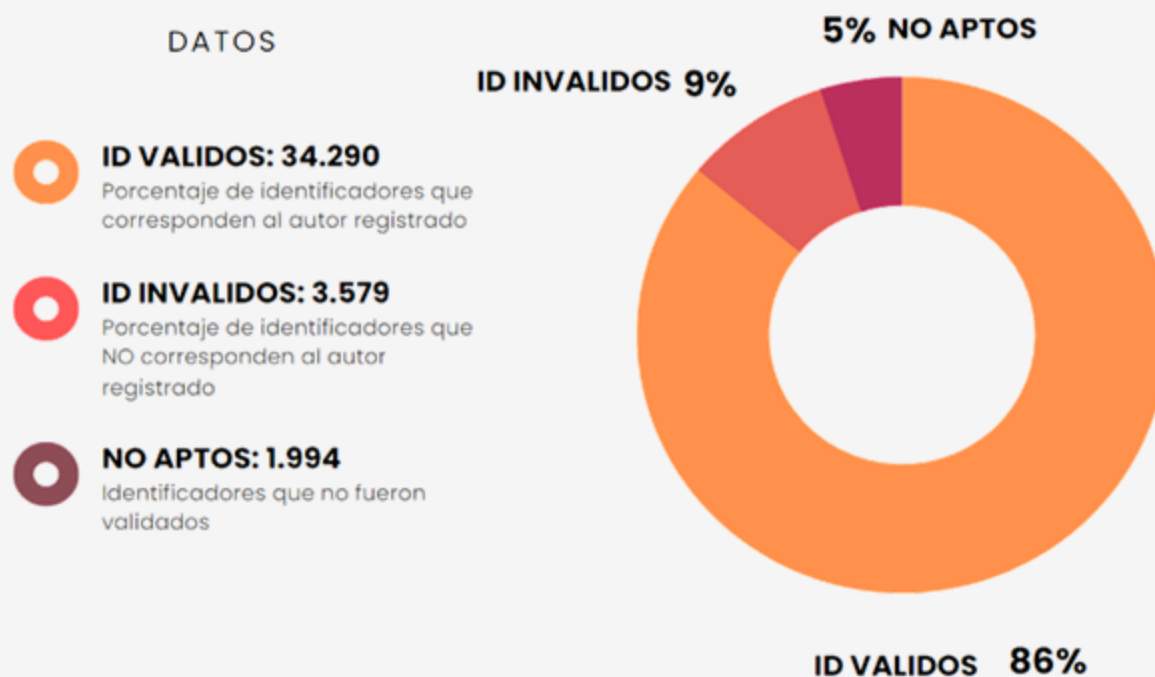
Uso de JSON por ser un formato estándar para el envío de datos mediante peticiones HTTP entre navegadores web y otras aplicaciones

Esquema Autor	Esquema Artículo
1 {	1 {
2 "id_cv": " ",	2 "marca_verificacion": true,
3 "Categoria": " ",	3 "tipo_articulo": " ",
4 "Nombre": " ",	4 "index": 0,
5 "Nombre en citas": " ",	5 "texto_original": " ",
6 "Nacionalidad": " ",	6 "nombre_articulo": " ",
7 "Sexo": " ",	7 "autores": {
8 "Par evaluador": true,	8 " ",
9 "Identificadores": {	9 " ",
10 "Autor ID (Scopus)": " ",	10 " ",
11 "Open Researcher and Contributor ID (ORCID)": " ",	11 },
12 },	12 "pais": " ",
13 "Redes": {	13 "nombre_revista": " ",
14 "Google Scholar": " ",	14 "issn_revista": " ",
15 "ResearchGate": " ",	15 "editorial_revista": " ",
16 },	16 "volumen": " ",
17 "Experiencia Laboral": {	17 "fasciculo": " ",
18 [	18 "pagina_inicio": " ",
19 "Lugar",	19 "pagina_fin": " ",
20 "Inicio",	20 "anio": " ",
21 "Fin"	21 "doi": " ",
22]	22 "palabras_claves": " ",
23 },	23 "sectores": " ",
24 "Articulos": {	24 "reconocimientos": " "
25 {	25 }
26 {	
27 {	
28 }	
29 }	

Implementación de la Herramienta

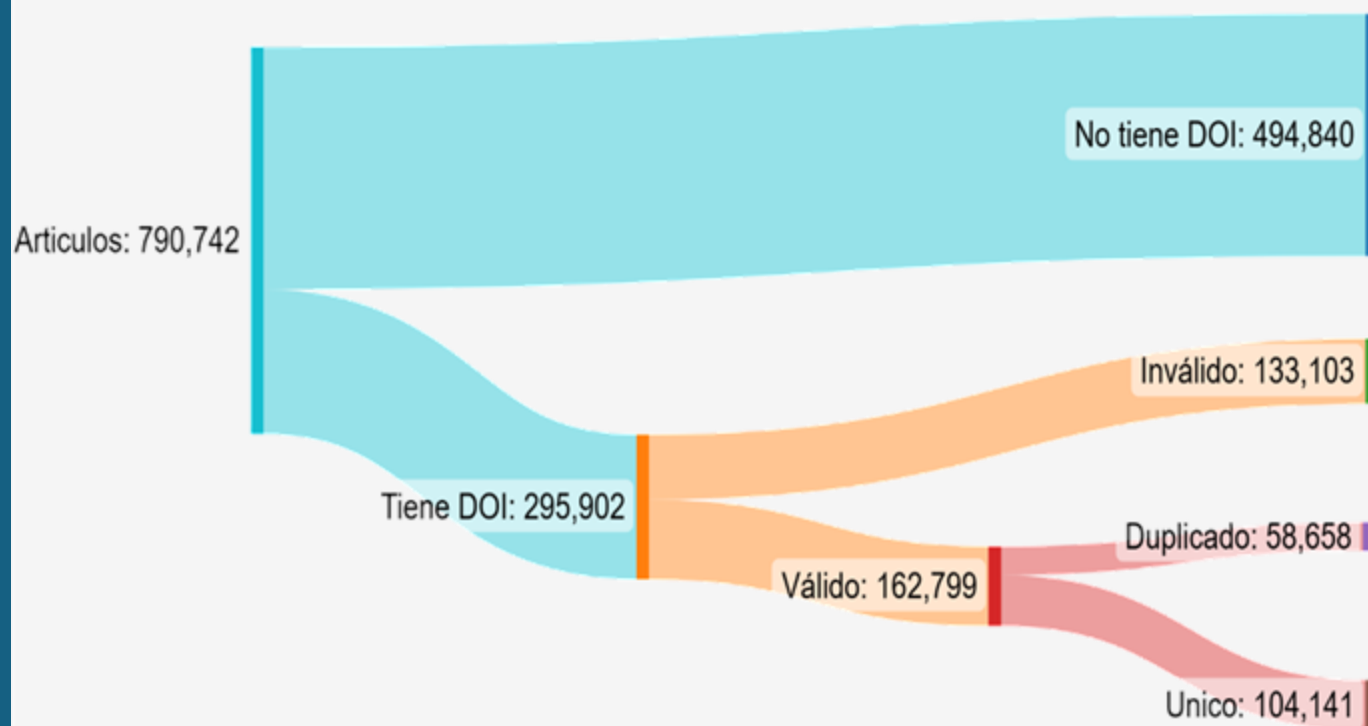


Verificación fiabilidad de los datos



- 39,792 registros conforme al método de filtrado propuesto, partiendo de un conjunto inicial de 56,702
- De los analizados 1,994 no cumplían con el formato correcto y 34,290 (86% del total) resultaron ser válidos tras el proceso de evaluación.

Verificación fiabilidad de los datos

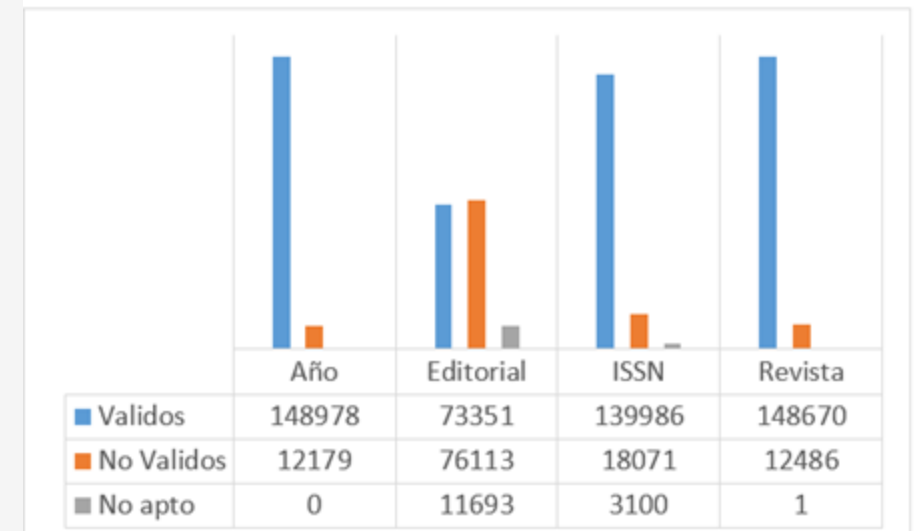


- 37,42% tiene un DOI asociado
- 20,60% presentan un DOI válido y verificado por medio del API de la DOI Foundation
- 13,15% contiene un DOI válido y único



Verificación fiabilidad de los datos (Artículos registrados en CvLAC vs Crossref)

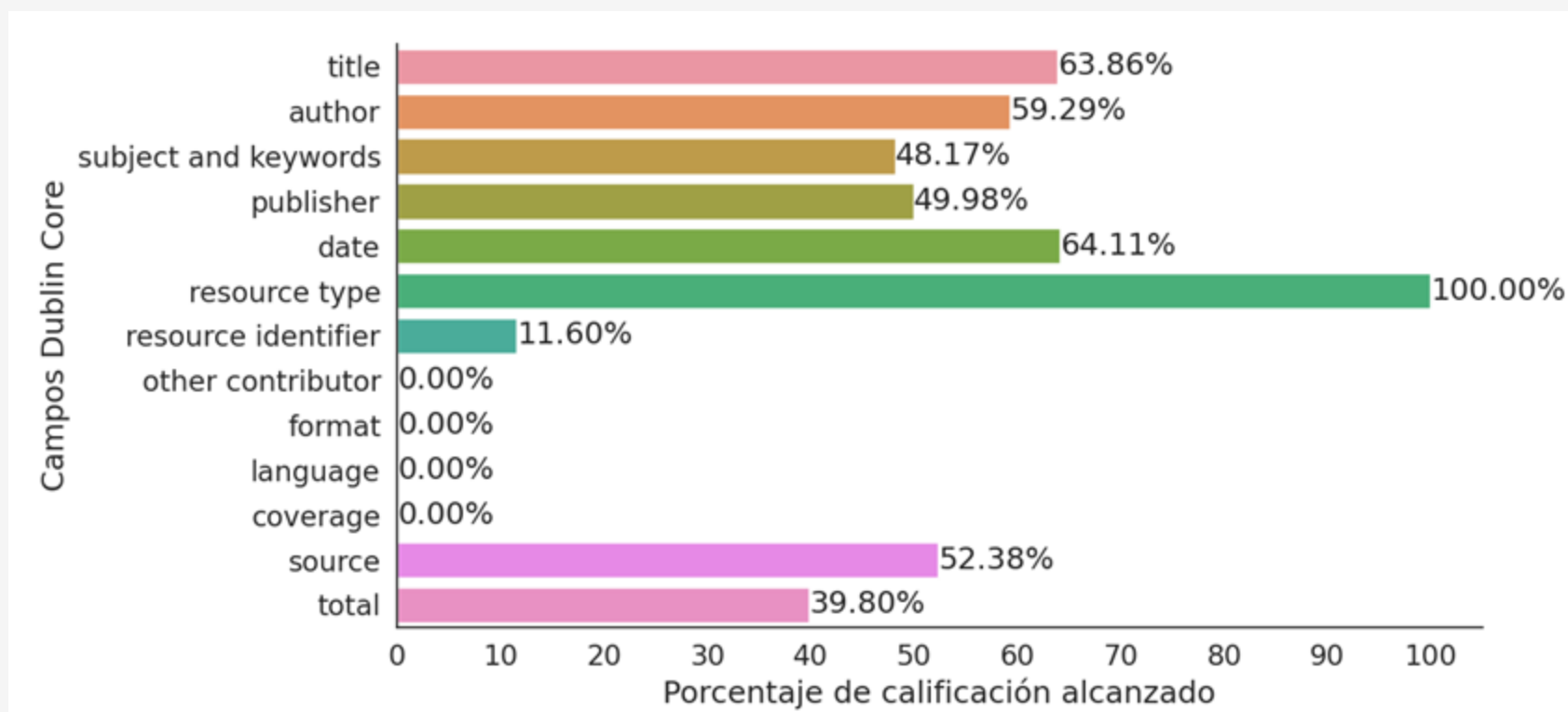
- Verificación de: Título, Autores, Año, Editorial, ISSN, Revista
- Títulos con gran concordancia: el 60% de artículos alcanza el 100% de concordancia en el título
- El 40% de artículos todos sus autores están validados



Evaluación de artículos



60.31% de los artículos obtuvieron una calificación entre 32 y 36 puntos de un total de 100 puntos



Conclusiones

Los errores más frecuentes detectados con el análisis de la información corresponden a:

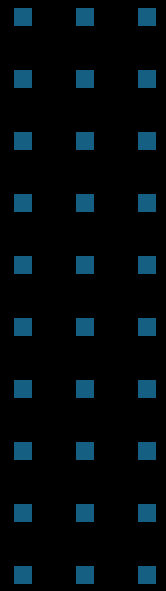
- i) registro de un número de autores diferente al que tiene el producto;
- ii) listado de artículos inexistentes; y
- iii) errores tipográficos o de mecanografía que no permiten validar el producto.

Errores que podrían ser, en parte, controlados mediante el uso de estándares internacionales de gestión de datos como Dublin Core, o incluso incluyendo validaciones simples en los campos del formulario de registro del producto.

Se debe propender por la interoperabilidad de plataformas y un intercambio de datos edficiente.

Referencias

- COVIDSurg Collaborative, & GlobalSurg Collaborative: SARS-CoV-2 infection and venous thromboembolism after surgery: An international prospective cohort study. *Anaesthesia*, 77(1), 28-39 (2022). <https://doi.org/10.1111/anae.15563>
- Dattolo, A., Corbatto, M.: Assisting researchers in bibliographic tasks: A new usable, real-time tool for analyzing bibliographies. *Journal of the Association for Information Science and Technology* 73(6), 757-776 (2022). <https://doi.org/10.1002/asi.24578>
- Galeano, D.F., Prada, L.C.: Diseño de un sistema inteligente para estimación de categorización de grupos de investigación a partir de lineamientos definidos por COLCIENCIAS (2020). <http://hdl.handle.net/11349/22537>
- Martínez, R., Rodríguez, R., Vera, P., Parkinson, C.: Análisis de técnicas de raspado de datos en la web aplicado al Portal del Estado Nacional Argentino. In: XXV Congreso Argentino de Ciencias de la Computación (2019).
- Mena-Chalco, J.P., Junior, R.M.C.: scriptLattes: An open-source knowledge extraction system from the Lattes platform. *Journal of the Brazilian Computer Society* 15(4), 31-39 (2009). <https://doi.org/10.1007/BF03194511>
- Ministerio de Ciencia, Tecnología e Innovación: Convocatoria Nacional para el Reconocimiento y Medición de Grupos de Investigación, Desarrollo Tecnológico o de innovación y para el reconocimiento de investigadores del Sistema Nacional de Ciencia (2020). https://minciencias.gov.co/sites/default/files/upload/convocatoria/anexo_1_-_documento_conceptual_2021.pdf
- Mosquera-Perdomo, A., Salazar Galindez, J., Ramirez-Gonzalez, G., Figueroa, C.: Software for the extraction of bibliographic information registered in CvLAC and GrupLAC applied in the Department of Cauca. *RIINN* 11(2), e3464 (2023). <https://doi.org/10.21897/rrii.3464>
- Patil, V.H., Bhavsar, S.A., Patil, A.H.: An efficient author information retrieval tool for bibliographic record analysis. *Journal of Intelligent and Fuzzy Systems* 39(1), 341-353 (2020). <https://doi.org/10.3233/JIFS-191289>
- Razzaq, S., Malik, A.K., Raza, B., Khattak, H.A., Moscoso Zegarra, G.W., Díaz Zelada, Y.: Research Collaboration Influence Analysis Using Dynamic Co-authorship and Citation Networks. *International Journal of Interactive Multimedia and Artificial Intelligence* 7(3), 103-116 (2022). <https://doi.org/10.9781/ijimai.2022.03.001>
- Riggio, G.: Indicadores bibliométricos de la actividad científica de la República Dominicana [Unpublished doctoral thesis]. Universidad Carlos III de Madrid (2017).
- Ruiz-Rosero, J., Ramirez-Gonzalez, G., Viveros-Delgado, J.: Software survey: ScientoPy, a scientometric tool for topics trend analysis in scientific publications. *Scientometrics* 121(2), 1165-1188 (2019). <https://doi.org/10.1007/s11192-019-03213-w>
- Valles-Coral, M.A.: Modelo de gestión de la investigación para incrementar la producción científica de los docentes universitarios del Perú. *Revista de Investigación, Desarrollo e Innovación* 10(1), 67-78 (2019). <https://doi.org/10.19053/20278306.v10.n1.2019.10012>



GRACIAS