Revista Facultad de Ingeniería

ISSN: 0121-1129

ISSN: 2357-5328

Universidad Pedagógica y Tecnológica de Colombia

González-Sanabria, Juan-Sebastián; Ramos-Corredor, Fabián-Nicolás; Amezquita-Becerra, Germán

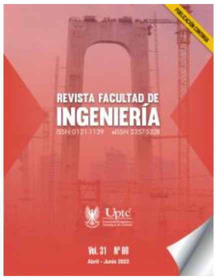
Automation Tool for Institutional Repositories Evaluation

Revista Facultad de Ingeniería, vol. 31, no. 61, e201, 2022, July-September

Universidad Pedagógica y Tecnológica de Colombia

DOI: https://doi.org/10.19053/01211129.v31.n61.2022.14724

Revista Facultad de Ingeniería
Journal Homepage: https://revistas.uptc.edu.co/index.php/ingenieria



Automation Tool for Institutional Repositories Evaluation

Juan-Sebastián González-Sanabria[1]
Fabián-Nicolás Ramos-Corredor[2]
Germán Amezquita-Becerra[3]

Received: April 18, 2022 **Accepted:** August 02, 2022 **Published:** August 08, 2022

Citation:
J.-S. González-Sanabria, F.-N. Ramos-Corredor, G. Amezquita-Becerra, "Automation Tool for Institutional Repositories Evaluation," *Revista Facultad de Ingeniería*, vol. 31 (61), e14724, 2022. https://doi.org/10.19053/01211129.v31.n61.2022.14724

Abstract

The rise of digital repositories has framed a significant advance in access to academic and scientific knowledge, increasing its impact due to greater reach and lower cost. However, these platforms are a new topic that initially did not have standards or models to carry out their implementation and operation, which is why there were inconsistencies between repositories on issues such as interoperability, digital preservation, among others. Due to the lack of standardization and the exponential increase in the number of repositories, different organizations and

[1] M. Sc. Universidad Pedagógica y Tecnológica de Colombia (Tunja-Boyacá, Colombia). juansebastian.gonzalez@uptc.edu.co. ORCID: 0000-0002-1024-6077
[2] Universidad Pedagógica y Tecnológica de Colombia (Tunja-Boyacá, Colombia). fabian.ramos01@uptc.edu.co
[3] M. Sc. Universidad Pedagógica y Tecnológica de Colombia (Tunja-Boyacá, Colombia). german.amezquita@uptc.edu.co. ORCID: 0000-0002-9001-1736

researchers made multiple proposals to standardize the processes and characteristics of these platforms. The proposals materialized in models, such as the Dublin Core and DataCite metadata schemes, and in guides for the evaluation and implementation of repositories, such as the "Guide for the evaluation of institutional research repositories" by RECOLECTA or the DINI certificate (Deutsche Initiative für Netzwerk Information). The latter aim to evaluate the platforms in their entirety, including 8 sections with a total of 87 elements. Therefore, in this research an application was developed to automate the evaluation of repositories, automating processes that improve educational work using computer tools and their integration.

Keywords: evaluation; institutional repositories; process automation; software development.

Herramienta para la automatización de la evaluación de repositorios institucionales

Resumen

El auge de los repositorios digitales ha enmarcado un avance significativo en el acceso al conocimiento académico y científico, aumentando su impacto debido a un mayor alcance y un menor costo. Sin embargo, estas plataformas son un tema novedoso que en un principio no contó con estándares o modelos para llevar a cabo su implementación y funcionamiento, por lo cual se presentaron inconsistencias entre repositorios en temas como interoperabilidad, preservación digital, entre otros. A causa de la falta de normalización y el incremento exponencial de la cantidad de repositorios, diferentes organizaciones e investigadores realizaron múltiples propuestas para estandarizar los procesos y características de dichas plataformas. Las propuestas se materializaron en modelos, como los esquemas de metadatos Dublin Core y DataCite, y en guías para la evaluación e implementación de repositorios, como la “Guía para la evaluación de repositorios institucionales de investigación” de RECOLECTA o el certificado DINI (Deutsche Initiative für Netzwerk Information). Estas últimas pretenden evaluar las plataformas en su totalidad incluyendo 8 apartados con un total de 87 elementos. Por lo anterior, en esta investigación se desarrolló un aplicativo para la automatización de la evaluación de repositorios, automatizando procesos que mejorar el quehacer educativo mediante el uso de herramientas informáticas y la integración de estas.

Palabras clave: automatización de proceso; desarrollo de software; evaluación; repositorios institucionales.

Ferramenta para automatizar a avaliação de repositórios institucionais

Resumo

A ascensão dos repositórios digitais tem enquadrado um avanço significativo no acesso ao conhecimento acadêmico e científico, aumentando seu impacto devido ao maior alcance e menor custo. No entanto, essas plataformas são um tema novo que inicialmente não possuía padrões ou modelos para realizar sua implementação e operação, razão pela qual havia inconsistências entre os repositórios em questões como interoperabilidade, preservação digital, entre outras. Devido à falta de padronização e ao aumento exponencial do número de repositórios, diferentes organizações e pesquisadores fizeram várias propostas para padronizar os processos e características dessas plataformas. As propostas materializadas em modelos, como os esquemas de metadados Dublin Core e DataCite, e em guias de avaliação e implementação de repositórios, como o "Guia para a avaliação de repositórios institucionais de investigação" da RECOLECTA ou o certificado DINI (Deutsche Initiative für Informações Netzwerk). Estes últimos visam avaliar as plataformas na sua totalidade, incluindo 8 secções com um total de 87 elementos. Portanto, nesta pesquisa foi desenvolvido um aplicativo para automatizar a avaliação de repositórios, automatizando processos que melhoram o trabalho educacional por meio do uso de ferramentas computacionais e sua integração.

Palavras-chave: automação de processos; avaliação; desenvolvimento de software; repositórios institucionais.

I. INTRODUCTION

A repository is defined as a “website that collects, preserves, and disseminates the academic production of an institution (or a scientific discipline), allowing access to the digital objects it contains and their metadata” [1]. Moreover, digital repositories should have four characteristics: the self-archiving, which involves the creator, owner or a third party submitting the content to the platform; the interoperability, which includes the use of standardized processes that will allow communication with other repositories through OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting); the open access to the full text; and the long term preservation [1].

Institutional repositories are one of the digital platforms available for disseminating knowledge nowadays, and its use is increasing compared to other options, such as journals. This can be due to different open access initiatives that have been developed in the last century, such as one of the most important ones called Budapest Open Access Initiative [2]. This initiative, also called BOAI, encourages those who sign it to promote and foster the Internet as a free tool for the exposition and sharing of academic and scientific information.

Therefore, digital repositories have become essential platforms thanks to the “open access” movement since they are the ideal medium for knowledge dissemination. This is reflected in the exponential increase of repositories on a global scale. According to OpenDOAR (Directory of Open Access Repositories) [4], in August 2006, there were 501 digital repositories registered, and in November 2021, there were 5778.

Nevertheless, due to a lack of normalization of repositories in terms of how they should operate and what they should include, the implementation of the repositories was carried out without following guidelines or standards, resulting in a problem for the users of these systems. Thus, various organizations, institutions and researchers have developed models or guidelines with information that allows assessing the quality of the repositories. Those criteria are focused on assessing specific topics, such as usability or the repository’s design. They can also include a more general perspective, where elements such as policies and interoperability are considered.

One of the first initiatives was the group DINI (German Initiative for Network Information) [5], which defined and normalized the "DINI Certificate for Document and Publication Services" [6]. The certification considers the following criteria: Visibility of the service, policies, advisory services for authors and publishing houses, legal aspects, information security, indexing and interfaces, access statistics, and long-term availability.

Its recommendations should be considered, even though they are not mandatory, since this certification is constantly updated, and recommendations may become requirements in future versions. The DINI and the certificate were created primarily to improve university platforms for knowledge dissemination in Germany, which is why it is not well-known outside of that country.

In the same line, the DRIVER project (Digital Repository Infrastructure Vision for European Research), which joined OpenAIRE, defines guidelines focused on the communication between repositories to establish a compatibility and interoperability standard that allows integrating the repositories with other platforms, such as harvesters. The criteria can be divided into two layers: syntactic (use of OAI-PMH and OAI_DC) and semantic (use of vocabulary) [7]. Moreover, the guidelines proposed include the sections: a) Textual resources, b) Metadata (Dublin Core), and c) OAI-PMH implementation.

Like DINI's proposal, each section of the model comes with some characteristics or elements that are mandatory and others that are recommended. Essentially, the DRIVER's guidelines intend to establish a common vocabulary as a basis for the OAI-PMH's and Dublin Core's data to guarantee interoperability among repositories.

OpenAIRE (Open Access Infrastructure for Research in Europe) published in 2010 the "Guidelines for Literature Repository Managers" [8], which are based on the DRIVER guidelines and provide directions to digital repository administrators so they can define and implement data management policies. This guide describes each metadata that should be included in the documents of the repositories by establishing a definition and some requirements and recommendations for each element. In addition, the most recent version (v4, 2018) includes three different metadata schemas, which are: DC (Dublin Core), DataCite [9], and OAI (Open Archive Initiative).

Such metadata schemas are used since repositories implement one schema or the other depending on their needs or requirements because there is not an international standard. The OpenAIRE guidelines include 32 attributes in total, which means that by using certain schemas, the recommendations cannot be followed, such as Dublin Core that only has 15 attributes. However, it is possible to comply with the guideline entirely by complementing several schemas' attributes or implementing a more complete schema regarding OpenAIRE guide.

Another initiative is the RECOLECTA project (or Open Science Harvester) which emerged in 2007 from the collaboration between the Spanish Foundation for Science and Technology (FECYT) and the University Libraries Network (REBIUN). This project created the "Guide for Research Institutional Repositories Evaluation" [3], which is defined as a tool for the self-evaluation of repositories. The guide includes the following criteria:

- Visibility
- Policies
- Legal aspects
- Publication's descriptive metadata
- Metadata interoperability and access to content
- Logs and statistics
- Data security, authenticity, and integrity
- Value-added services and functionalities

In its fourth edition, this guide includes novelties such as the addition of: vocabulary for certain metadata created by COAR (Confederation of Open Access Repositories), the DataCire metadata schema, and levels defined by the National Digital Stewardship Alliances (NDSA) for the digital preservation of content. OpenAIRE's guidelines are used as a reference for metadata validation. Therefore, with the support of the institutions that propose it and the inclusion of other well-known guidelines and standards, this evaluation proposal is one of the most recognized and accepted by the community.

Finally, despite having more recent, complete, and recognized evaluation models, the software available for completing the evaluations is scarce. In Latin America, an interdisciplinary group developed an open source web application called "dPyx - Self-Assessment Tool for Academic and Scientific Information Systems [10], to provide the information platforms administrators with a set of indicators that allow them to evaluate their systems.

The dPyx indicators are based on good practices, documentation and international standards such as ISO 16363:2012, OpenAIRE, WCAG, and OAIS. In this platform, there are eight criteria or sections [10]:

- *Governance*: mandates, policies, resources, funding.
- *Maintenance and development*: guidelines, collections, roles, processes.
- *Accessibility*: platforms, speed, formats, permanence.
- *Software*: stability, updates, protocols, security.
- *Hardware*: updates, maintenance, connectivity.
- *Digital preservation*: formats, licences, processes, standards.
- *Positioning and visibility*: indexes, search engines, directories, metadata, interoperability.
- *Ethics and integrity*: good practices, transparency, FAIR.

This platform has three roles: users (repository/journal), administrator, and evaluator. Additionally, there are evaluation models for scientific journals and digital repositories. The system's operation begins with the request to create an administrator-type member who defines the evaluators and users. Then, the users are those people representing a journal or repository and that are in charge of completing the evaluation with a questionnaire. Finally, the evaluators validate that the information registered by the users is correct, in addition to making observations and approving or rejecting the qualification criteria [10].

The starting point was the definition of a new model composed of the following criteria: visibility, policies, legal aspects, metadata, interoperability, logs and statistics, security, authenticity and integrity of the data. The Alicia Guide 2.0 [11] was the reference for selecting these criteria. This guide adopts the OpenAIRE guidelines, the Dublin Core data schema, and criteria from other guidelines such as the DINI certificate.

Considering what was previously mentioned, this research aims at developing an application for the automation of repository evaluation, and it is executed in three stages: the definition of the evaluation criteria, the establishment of tools and services for making the automation, and the development of the software. The application intends to provide the community with a tool that will help in the process of improving open access to science.

II. Methodology

This research was developed in three stages: the systematic literature review, where we searched for proposals for the evaluation of repositories in the Google Scholar, Collector of Open Science (RECOLECTA), and LA Referencia harvesters. Then, we discarded the proposals already considered in more recent guidelines or that included aspects unrelated to using repositories as platforms, such as "La Accesibilidad Web en los Repositorios Institucionales. La UOC a examen" [12]. This guide evaluates the repository with a norm about accessible design that includes all types of web platforms. The filtered information consolidated an evaluation guide composed of the criteria considered in other guides.

In the second part, the viability of the automation of each of the elements was determined considering diverse aspects, such as the level of access (public or private) required to retrieve the information or the location of the data in terms of being normalized within the repository or outside of it. Subsequently, the tools and external services necessary to obtain the information that answered the automated items were identified.

In the third and last part, the evaluation prototype was developed from the guide established in the first section and the tools and services of the second section. The prototype was made with the Scrum framework with four sprints or iterations. After completing each, there was a presentation and revision of the developed criteria to determine if the implementation was adequate in accordance with the results obtained. Moreover, after creating the prototype, there was a validation and general verification process of the results. This was done to review the veracity of the information provided in the automated elements and validate if any of the automated elements was faulty since this type of application depends on a specific external service or web page.

III. RESULTS

In the documentary revision, we found three guidelines with international recognition: RECOLECTA [3], DINI [6], and ALICIA [11]. The last one is based on the first two, mainly on RECOLECTA, but it adds definitions to the metadata and legal aspects. Moreover, it includes a section called "IT Support" focused on the implementation and maintenance of the repository.

On the other hand, the guides from RECOLECTA and DINI have similar criteria; the only difference is the number of elements in each of the criteria. RECOLECTA has a higher number of elements, which represents greater detail in the evaluation. For example, in the metadata section of the DINI's guide, only basic elements and characteristics of the DC schema are reviewed, while RECOLECTA considers three different schemas and has a vocabulary for some of the elements.

Considering the aforementioned, RECOLECTA was used as a reference for the automation guide. The evaluation guide considers eight criteria:

- Visibility: This section mentions the aspects that give the repository greater recognition in a quest to publicize the platform and its content.

- Presence in international directories: OpenDOAR, ROAR, OAI Data Providers, re3data
- Presence in international harvesters: LA Referencia, OpenAIRE, Google Scholar, CORE, BASE
- Presence in national harvesters
- Use of a normalized name of the IR in directories and harvesters
- Use of a secure (https) and friendly (name of the IR) URL
- Availability of documents in open access
- Creation of initiatives to promote the visibility of the repository within the same institution

- **Policies:** It concerns the organization and governance of the repository to know its state and progress in terms of the definition of guidelines, norms, activities and processes.

    - Implementation of an open access policy
    - Adherence to the Budapest Declaration, one of the bases of the open access movement
    - Creation of an action policy of the IR (unified public document)
    - The information about the policy must be distributed on the IR web page
    - Established mission and objectives of the IR
    - Information on who can deposit, what can be deposited and in which formats
    - Information on how the contents are preserved
    - Information on the reuse of metadata
    - The contact information must be visible

- **Legal aspects:** Description of the management of copyrights

    - The authors must acknowledge that they are not violating any intellectual property right
    - The authors must sign an authorization for the distribution of their work
    - It should be stated how the authors can know if their work can be deposited in accordance with the editorial policy (Sherpa/Romeo, Dulcinea)
    - Including the copyright in the metadata of each resource
    - Including the copyright in each resource

- **Metadata:** The metadata are structured or semi-structured information that describes the content, quality, conditions, history, availability, and other characteristics of the documents, including data such as authors, date of publication, references, language, type, among others. This section defines the characteristics, format and vocabulary of the metadata that each document in the repository must include

    - Uses Dublin Core (DC) metadata schema • Includes author identifiers (ORCID, IraLIS)
    - Includes the following fields:
    - Author (dc:creator)

- Title (dc:title)
- Type of result of the research (dc:type)
- Resource version (dc:type)
- Date of publication (dc:date)
- Copyright (dc:rights) • Includes the following fields:
- Description (dc:description)
- Format (dc:format)
- Language (dc:language)
- Identifier (dc:identifier)
- Subject/descriptors/keywords (dc:subject)
- Contribution (dc:contributor)
- Funding reference (dc:relation)
- Publisher (dc:publisher)
- The field for access rights follows the established vocabulary (closedAccess, embargoedAccess, openAccess, restrictedAccess)
- The date of publication field follows the established format (ISO 8601 – YYYY-MM-DD, YYYY-MM-DDTHH:MM:SSZ)
- The language field follows the established vocabulary (ISO 639-1, 639-2 and 639-3, code zxx)
- The type of result of the research field contains only one occurrence
- The type of result of the research field is assigned following the vocabulary of resource type by COAR (Annex 1)
- The format field is assigned following the established vocabulary (Annex 2)
- The resource version field contains only one occurrence
- The resource version field follows the COAR vocabulary (draft, submittedVersion, acceptedVersion, publishedVersion, updatedVersion)
- A normalized classification system is implemented (availability of one or several normalized classification systems such as CDU, JEL, UNESCO)
- A technical or preservation metadata schema is used
- The repository develops some sort of activity of metadata curation
- The metadata are exported in a format different from DC

- **Interoperability:** Declaration of processes and characteristics of the services of content extraction of the platform
    - Harvested by LA Referencia-OpenAIRE
    - The metadata are provided through the OAI-PMH protocol
    - The deleted records are marked
    - The life span of the resumption token is of at least twenty-four hours
    - The email of the repository's administrator is available on the tag AdminEmail within the response to an Identify order
    - There is a Description declaration in the response to an Identify order
    - The delivery of records through the OAI-PMH protocol is progressive by batches
    - The size of the batches for the delivery of records is within the range of 100-500 records
    - The format of the date in the Identify order matches the field datestamp of the records
    - Contemplates integration with other information systems in the institution
    - Includes <meta…> tags in the HTML heading https://scholar.google.com/intl/es/scholar/inclusion.html#indexing)
    - The repository supports other protocols and APIs to share metadata or content
    - Widespread use of persistent identifiers (DOI, Handle, URN, ORCID)
- **Security** Corresponds to the evaluation of the practices and strategies used by the administrators of the repositories to maintain the integrity and reliability of the information where it is stored and the processes where it is transferred.
    - The IR web page informs about the creation of security copies
      The IR web page informs about the execution of checksums
    - There are at least three copies of the records (metadata and files), and at least one of them is located in a different geographic location
    - Identification, control and validation of formats (JHOVE, DROID, Xena)

- **Statistics and logs:** This criterion reviews the information on the access and use of the platform from the users' side on a general level and for each document. Additionally, the way this information is stored is verified.
    - Availability of public statistics of the IR in general
    - Availability of public statistics of each document in the IR
    - The logs of the web server where the repository is hosted are permanently archived
    - The COUNTER standard is used
- **Value-added services:** The novel, different and value-adding characteristics are essential to improving the user's perception. Therefore, this item evaluates if the repository includes those differential services that can position it over other platforms.
    - Social networks are used to share each document (Twitter, Facebook, LinkedIn)
    - Integrates bibliographic managers (Zotero, Mendeley)
    - Visualizes and exports the metadata in different schemas (METS, PREMIS, RDF, JSON, MARC, BibTeX)
    - Alert services (RSS) are available
    - There are author profiles
    - The repository offers metrics based on citations
    - Next-generation metrics (such as the h-index)
    - Offers external tools and services

After establishing the elements of the guide, it was defined that it was not possible to automate them all since the information of some items was obtained through webs external to the repository or was not published on the Internet. The issue with external web pages is that they tend to have a different structure and organization. Therefore, the information is not always found in the same places. In the case of unpublished information, the required data are only known by the system's administrators.

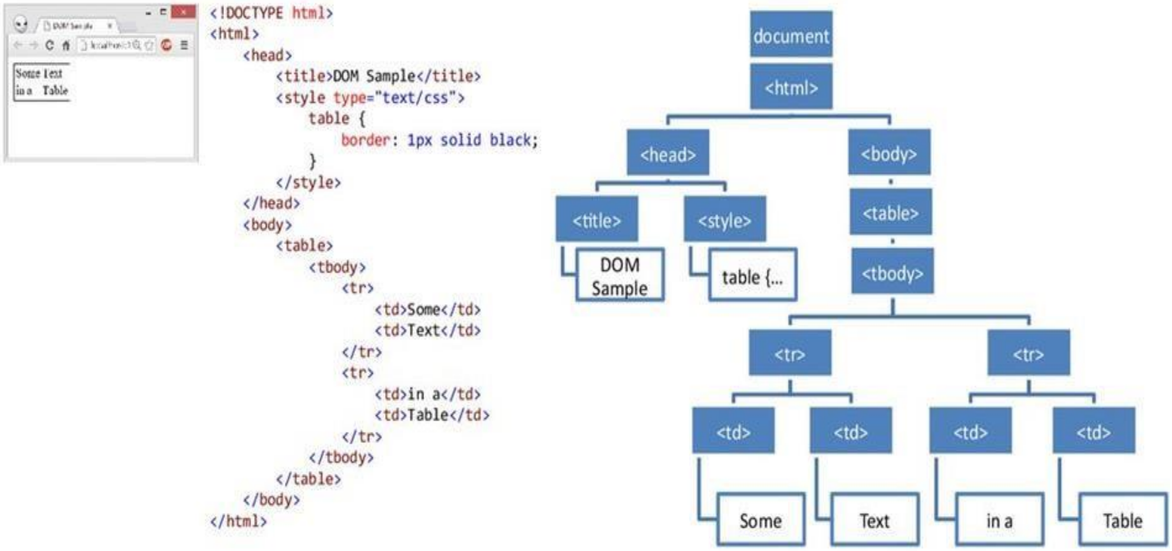
Some examples of the presented situations are: "Information on who can deposit, what can be deposited and in which formats" (policies), where the information is on the network, but its location is not normalized. Moreover, “The logs of the web server where the repository is hosted are permanently archived”, where the data are specific to the operation of the application.

After determining the items that could not be automated, the tools and services for the rest of the elements were established. First, it was determined that for the items related to OpenDOAR and Google Scholar, it was possible to use the APIs that were proprietary or external to these platforms, specifically Sherpa API and Scale SERP for the first and the second, respectively. For the rest of the harvesters and directories, the web scraping technique was used since the information required was included in the HTML of the platforms’ web pages.

For the ROAR and OAI Data Providers directories, a different process was implemented since ROAR did not allow to make requests through web scraping, and OAI Data Providers had a static web page that was not updated frequently, which made inefficient downloading the web page for each evaluation to obtain the same information most of the time. To solve this, we created a script that downloads the data from OAI Data Providers and uploads an XML file (obtained from ROAR’s web page) to create then a relational database with the information from both platforms.

Finally, we used web scraping for everything related to metadata and attributes of the documents or published works in the repositories by searching in the HTML document the tags corresponding to the desired elements, for example, <meta> for the DC schema data. However, the URLs of the documents depend on the findings of the directories and harvesters since each repository organizes the information according to its needs and disciplines. Thus, it was not feasible to standardize the retrieval of the links directly from the search in the repositories. Furthermore, this dependency means that the items related to the documents will not be evaluated if no information is found in the first part.

For the web scraping, we used the Document Object Model (DOM) parsing method, where the HTML of the web page is converted into a DOM tree from which the data can be obtained by searching their attributes, tags or relations. One example of this is shown in Figure 1.



Some Text
in a Table

```
<!DOCTYPE html>
<html>
    <head>
        <title>DOM Sample</title>
        <style type="text/css">
            table {
                border: 1px solid black;
            }
        </style>
    </head>
    <body>
        <table>
            <tbody>
                <tr>
                    <td>Some</td>
                    <td>Text</td>
                </tr>
                <tr>
                    <td>in a</td>
                    <td>Table</td>
                </tr>
            </tbody>
        </table>
    </body>
</html>
```

Fig. 1. Example of Document Object Model (DOM) parsing extracted from the presentation Module 7: Accessing DOM with JavaScript [13].

The architecture of the tool's communication with the application is presented in Figure 2.

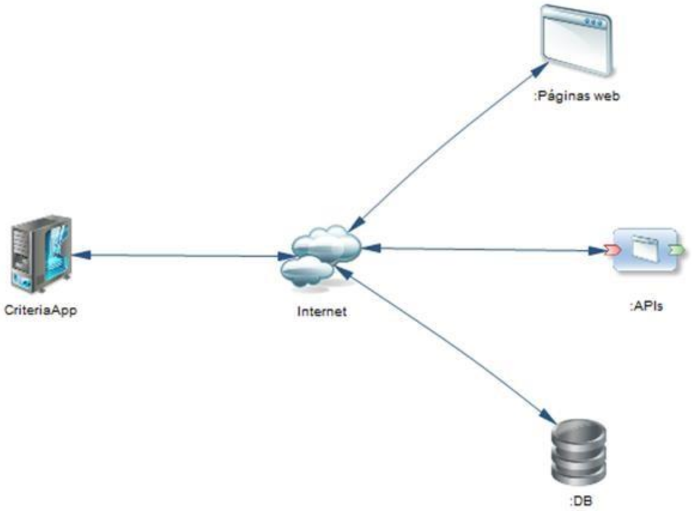


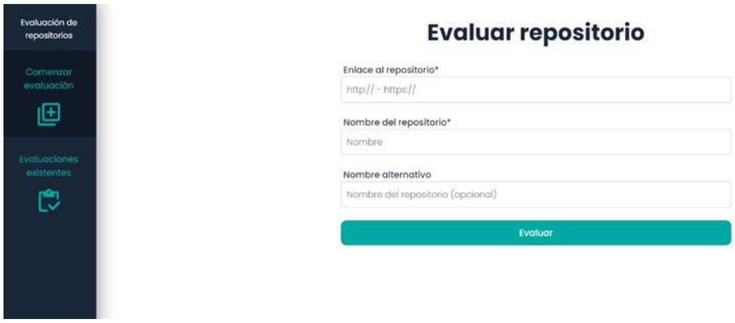
Fig 2. Architecture of the application's tools.

For the development of the application, a client-server structure was selected to have different users simultaneously. Then, two frameworks were selected for the software: React for the client and Flask for the server. This selection was made due to previous experiences. Additionally, Flask offered an advantage regarding web

DOI: https://doi.org/10.19053/01211129.v31.n61.2022.14724

scraping because it is based on Python. This language has one of the most complete libraries for the development of this technique. This section describes the screens of the systems with their functionalities.

On the main page of the application (https://criteria-front.herokuapp.com/home) there are two main components: a form to start the evaluation of the repository and a menu with the options Start Evaluation and Previous Evaluations (Figure 3).



Evaluar repositorio

Enlace al repositorio*
http:// - https://

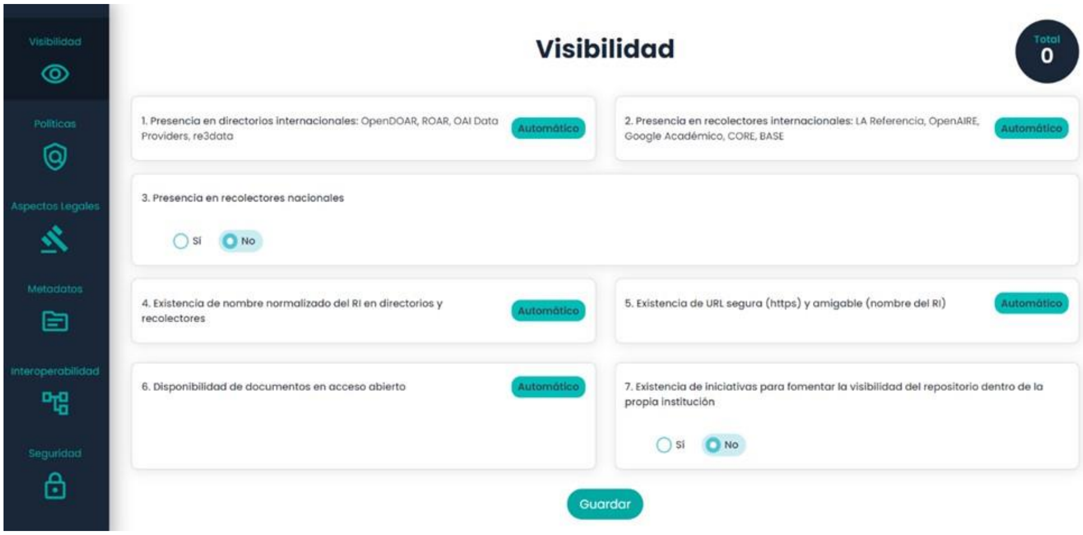
Nombre del repositorio*
Nombre

Nombre alternativo
Nombre del repositorio (opcional)

Fig. 3. Homepage.

The data required in the form are: link to the repository, name of the repository and alternative name (optional). The link is required because certain items validate that the URL matches the repository's URL. The two names are used for searching in international harvesters and directories where the URLs of the documents to be evaluated are obtained. Additionally, it is worth mentioning that the two names are requested for the lack of normalization of names on different platforms.

Then, a series of data is requested for each characteristic. For example, for Visibility, there is a menu with diverse criteria and some cards with the different items corresponding to visibility. In the automatic elements, there is a tag stating this, and those that are not automatic have the option of choosing whether it complies with the condition (Figure 4).



Visibilidad

1. Presencia en directorios internacionales: OpenDOAR, ROAR, OAI Data Providers, re3data — Automático

2. Presencia en recolectores internacionales: LA Referencia, OpenAIRE, Google Académico, CORE, BASE — Automático

3. Presencia en recolectores nacionales

○ Sí ● No

4. Existencia de nombre normalizado del RI en directorios y recolectores — Automático

5. Existencia de URL segura (https) y amigable (nombre del RI) — Automático

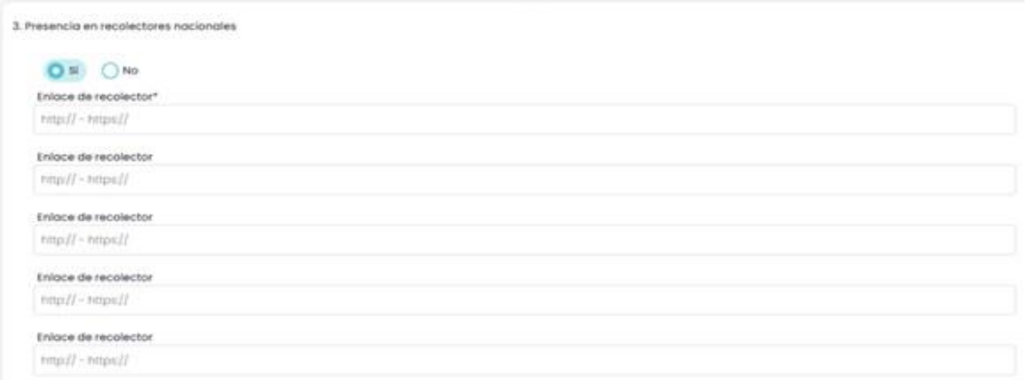
6. Disponibilidad de documentos en acceso abierto — Automático

7. Existencia de iniciativas para fomentar la visibilidad del repositorio dentro de la propia institución

○ Sí ● No

Fig. 4. Visibility page.

There is verification in the international harvesters and directories to make sure that one of the names found on these platforms has a similarity of at least 90% concerning those entered on the homepage, given that in some cases, they do not match entirely due to accents or connectors. In the item Presence in national harvesters, if the answer is affirmative, at least one URL of the five possible entries must be entered, as shown in Figure 5.



3. Presencia en recolectores nacionales

● Sí ○ No

Enlace de recolector*
http:// - https://

Enlace de recolector
http:// - https://

Enlace de recolector
http:// - https://

Enlace de recolector
http:// - https://

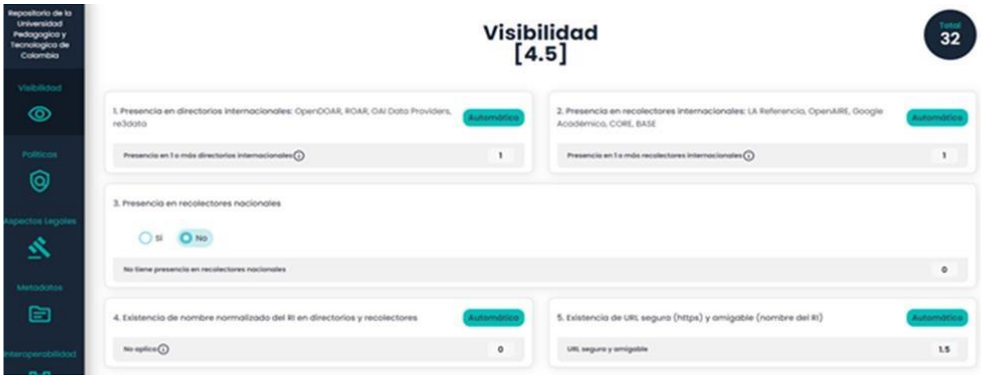
Enlace de recolector
http:// - https://

Fig. 5. Item presence in national harvesters.

The evaluation of the fourth element involves verifying if the names found in the directories and harvesters are the same. The fifth item has two parts: verifying if the repository's URL has https, and making sure that the URL is no longer than 40 characters and that any of them is a special character, which is defined as a friendly

URL. The result of the sixth element is obtained by determining if in all the documents obtained in the directories and harvesters the phrase openAccess is present in the metadata with the tag DC.rights. The last item only involves selecting Yes or No as applicable.

After entering the information, the Save button must be pressed, which sends the data to the server where the URLs are verified. If the URLs are correct, the result of the criteria is returned; otherwise, the errors are displayed. Figure 6 presents an example of the result for the criteria.



Visibilidad
[4.5]

1. Presencia en directorios internacionales: OpenDOAR, ROAR, OAI Data Providers, re3data

Presencia en 1 o más directorios internacionales 1

2. Presencia en recolectores internacionales: LA Referencia, OpenAIRE, Google Académico, CORE, BASE

Presencia en 1 o más recolectores internacionales 1

3. Presencia en recolectores nacionales

○ Si ● No

No tiene presencia en recolectores nacionales 0

4. Existencia de nombre normalizado del RI en directorios y recolectores

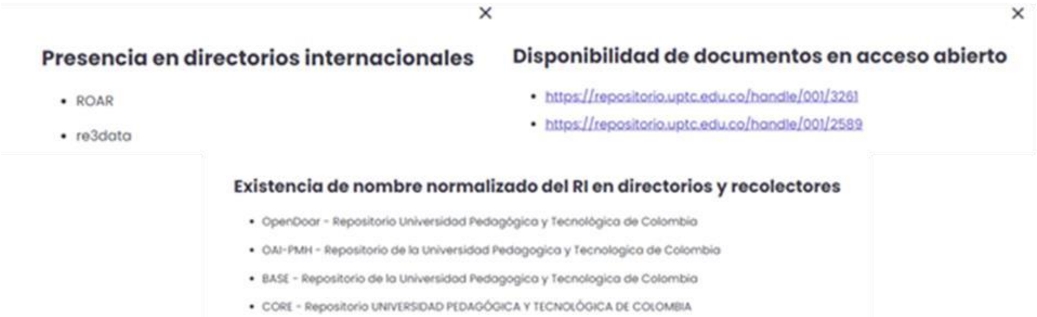
No aplica 0

5. Existencia de URL segura (https) y amigable (nombre del RI)

URL segura y amigable 1.5

Fig. 6. Result of the evaluation of the criteria Visibility.

The results are shown for each item, both by criteria and as a total, which is the sum of all the criteria evaluated. Additionally, for each item, a message, the score, and in some cases, a button opening a new window with additional information (which corresponds to the documents or platforms that did not comply with the requirements) are displayed. Some examples of the extra information are presented in Figure 7.



Presencia en directorios internacionales

- ROAR
- re3data

Disponibilidad de documentos en acceso abierto

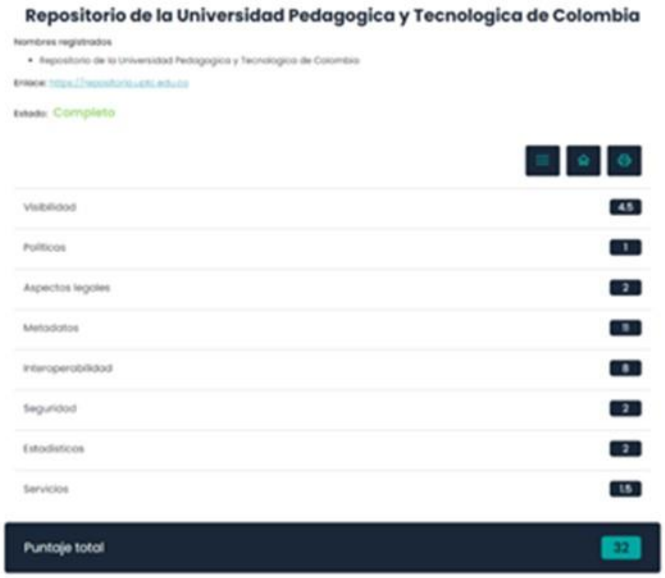
- https://repositorio.uptc.edu.co/handle/001/3261
- https://repositorio.uptc.edu.co/handle/001/2589

Existencia de nombre normalizado del RI en directorios y recolectores

- OpenDoar - Repositorio Universidad Pedagógica y Tecnológica de Colombia
- OAI-PMH - Repositorio de la Universidad Pedagogica y Tecnologica de Colombia
- BASE - Repositorio de la Universidad Pedagogica y Tecnologica de Colombia
- CORE - Repositorio UNIVERSIDAD PEDAGÓGICA Y TECNOLÓGICA DE COLOMBIA

Fig. 7. Example of additional information in the evaluation.

Moreover, Figure 8 presents the result of the evaluation process of the repository of Universidad Pedagódica y Tecnológica de Colombia.



Repositorio de la Universidad Pedagogica y Tecnologica de Colombia

Nombres registrados

- Repositorio de la Universidad Pedagogica y Tecnologica de Colombia

Enlace: https://repositorio.uptc.edu.co

Estado: Completo

Visibilidad	4.5
Politicas	1
Aspectos legales	2
Metadatos	11
Interoperabilidad	8
Seguridad	2
Estadisticas	2
Servicios	1.5
Puntaje total	32

Fig. 8. Screen of the summary of the evaluation.

Additionally, this page has three buttons that redirect to the detail of the result of each criterion and the homepage. There is also the option of downloading the result of the evaluation.

IV. Conclusions

The evaluation of repositories is a topic that has gained relevance within the academic and scientific community with the rise and growth of these platforms. Therefore, different institutions, organizations and researchers have proposed guidelines for standardizing repositories. Over time, some guidelines and documents with more elements have consolidated in the field. One example of this is the “DINI Certificate for Document and Publication Services” and the “Guide for the Evaluation of Institutional Research Repositories”.

Given the intended focus of the system developed in this research, the RECOLECTA guide was considered to define the evaluation criteria since it is the most widely known in the scientific community and is constantly updated. For example, in 2021, new metadata schemas and vocabularies were added.

Three elements were chosen in selecting the external tools and services: APIs, web scraping and databases. The first is in the case of some harvesters and directories that have this service. The second was mainly used for the sections where it was needed to evaluate the information in the documents of the repositories. The third was established for two directories that did not have an API, and their information did not vary frequently. Thus, it was more efficient to define a database with the information from these platforms and update it frequently.

Finally, the React and Flask frameworks were used to develop the application due to previous experiences. Moreover, Flask was used because it is based on Python, which has the most complete libraries for implementing web scraping. Although not all the evaluation items were automated, it is worth noting that the items that were automated were the most time-consuming. These were related to the documents, where it was necessary to review each of them, looking for the information and formats required in the guide.

In general, the application allowed to decrease the time required for the evaluation significantly. This was possible thanks to the standardization of the content of the documents in the repositories. This research is an example of how it is possible to automate processes or routines on previously established schemas and data by means of techniques such as web scraping.

AUTHORS' CONTRIBUTION

Juan-Sebastián González-Sanabria: Investigation, Methodology, Witing-review & editing.

Fabián-Nicolás Ramos-Corredor: Investigation, Methodology, Witing-orginal draft, Software.

Germán Amezquita-Becerra: Investigation, Validation; Witing-review & editing.

REFERENCES

[1] C. González Díaz, M. Iglesias García, M. Martín Llaguno, A. González Pacanowsky, *Antecedentes y estado de la cuestión sobre los Repositorios Institucionales de Contenido Educativo (RICE)*, Departamento de Comunicación y Psicología Social, 2015

[2] *Budapest Open Access Initiative*, 2002. https://www.budapestopenaccessinitiative.org/read

[3] C. Azorín, I. Bernal, J. Gómez Castaño, C. Guzmán Pérez, M. Losada Yáñez, R. Marín del Campo, F. J. Martínez Galindo, C. Martínez Pousa, J. C. Morillo Moreno, J. Prats Prat, *Guía para la evaluación de repositorios institucionales de investigación*, Fundación Española para la Ciencia y la Tecnología, 2021.

[4] Jisc, "*OpenDOAR Statistics*," Universidad de Nottingham, s. f. https://v2.sherpa.ac.uk/view/repository_visualisations/1.html

[5] AMH (Arbeitsgemeinschaft der Medieneinrichtungen an Hochschulen e. V.); dbv (Deutscher

[6] Bibliotheksverband e. V., Sektion 4: Wissenschaftliche Universalbibliotheken); ZKI (Zentren für Kommunikation und Informationsverarbeitung in Lehre und Forschung e. V.), *DINI - Deutsche Initiative für Netzwerkinformation*, s. f. https://dini.de

[7] Grupo de trabajo "Publicación Electrónica", *Certificado DINI Servicio de Documentación y Publicaciones*, Göttingen, Alemania, 2012.

[8] M. Vanderfeesten, F. Summann, S. Martin, *Directrices DRIVER 2.0*, 2008.

[9] OpenAIRE, *Directrices de OpenAIRE para administradores de repositorios de Literatura v4*. https://guiasopenaire4.readthedocs.io/es/latest/index.html

[10] DataCite Metadata Working Group, *DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs*, 2021.

[11] Escire, *dPyx - Herramienta de autoevaluación de sistemas de información académica y científica*. https://dpyx.site

[12] Red Nacional de Repositorios Digitales de Ciencia, Tecnología e Innovación de Acceso Abierto (RENARE), *Guía Alicia 2.0*, Lima, Perú, 2020.

[13] M. Á. Bolaños Asenjo, *La Accesibilidad Web en los Repositorios Intitucionales. La UOC a examen*, Universitat Oberta de Catalunya, 2012.

[14] SoftServe, *Module 7: Accessing DOM with JavaScript*. https://ppt-online.org/83335