INGENIERÍA E INNOVACIÓN
RIINN
Revista Científica Facultad de Ingeniería
SN: 2346-0474 (En línea)

Software for the extraction of bibliographic information registered in CvLAC and GrupLAC applied in the Department of Cauca

Software para la extracción de información bibliográfica registrada en CvLAC y GrupLAC aplicado en el Departamento del Cauca

* *Alexander Mosquera-Perdomo* *University of Cauca,*

Jarby Salazar-Galindez *University of Cauca,*

Gustavo Ramirez-Gonzalez *University of Cauca,*

Cristhian Figueroa *University of Cauca,*

RESUMEN

The access and use of bibliographic information from scientific activity in Colombian territories is of vital importance to know the impact, productivity, collaboration and evolution of regional and national research. However, this information is presented in a static and unstructured manner in the official national platforms available for this purpose, thus limiting its analysis for the strategic decision making by the actors in the science, technology and innovation environment. This research proposes a tool for extracting, structuring and persisting data registered by researchers and research groups in the Colombian territories. It is aimed to improve the efficiency and ease of access to bibliographic data in order to support stakeholders. As methodology, an adaptation of the Project Management Body of Knowledge was used, defining phases of analysis, design, coding and evaluation during the development of the software. The results obtained allowed to successfully verify and validate the system, to point out both disparities and contrasts with respect to the statistics offered by the Ministry of Science, Technology and Innovation, to acquire multiple organized data sets about scientific activity in Cauca, and to offer a tool for obtaining more data sets from other departments of the country. Finally, it is important to emphasize that this tool seeks to contribute to the incentive, strengthening and reliability of future analysis and conclusions in research work or decision-making by individuals and entities related to the subject.

PALABRAS CLAVES: Data extraction, Bibliography, Research, CvLAC, GrupLAC.

1. *University of Cauca, Faculty of Electronic and Telecommunication Engineering, Popayán, Colombia, edisonm@unicauca.edu.co, https://orcid.org/0009-0002-0784-3046*
2. *University of Cauca, Faculty of Electronic and Telecommunication Engineering, Popayán, Colombia, dsalazarg@unicauca.edu.co, https://orcid.org/0009-0001-5793-0538*
3. *Telmatics Engineering Doctor of Philosophy, University of Cauca, Professor, Faculty of Electronic and Telecommunication Engineering, Popayán, Colombia, gramirez@unicauca.edu.co, https://orcid.org/0000-0002-1338-8820Doctor en Ciencias de la Educación, Docente /grupo INGLEX/Facultad de Educación*

ABSTRACT

The access and use of bibliographic information from scientific activity in Colombian territories is of vital importance to know the impact, productivity, collaboration and evolution of regional and national research. However, this information is presented in a static and unstructured manner in the official national platforms available for this purpose, thus limiting its analysis for the strategic decision making by the actors in the science, technology and innovation environment. This research proposes a tool for extracting, structuring and persisting data registered by researchers and research groups in the Colombian territories. It is aimed to improve the efficiency and ease of access to bibliographic data in order to support stakeholders. As methodology, an adaptation of the Project Management Body of Knowledge was used, defining phases of analysis, design, coding and evaluation during the development of the software. The results obtained allowed to successfully verify and validate the system, to point out both disparities and contrasts with respect to the statistics offered by the Ministry of Science, Technology and Innovation, to acquire multiple organized data sets about scientific activity in Cauca, and to offer a tool for obtaining more data sets from other departments of the country. Finally, it is important to emphasize that this tool seeks to contribute to the incentive, strengthening and reliability of future analysis and conclusions in research work or decision-making by individuals and entities related to the subject.

KEYWORDS: Data extraction, Bibliography, Research, CvLAC, GrupLAC.

*Universidad del Atlántico, Barranquilla, Colombia y rafaeloyaga@mail.uniatlantico.edu.co, 3223098858, Calle 69c35-27 Barranquilla, https://orcid.org/0000-0002-7830-9396**

4. Telmatics Engineering Doctor of Philosophy, University of Cauca, Professor, Faculty of Electronic and Telecommunication Engineering, Popayán, Colombia, cfigmart@unicauca.edu.co, https://orcid.org/0000-0001-6697-886X

Fecha recepción:
Fecha aceptación:

INTRODUCCIÓN

Scientific research is essential for the development and growth of a country and its regions. It has been demonstrated that research activity generates wealth, knowledge, welfare and innovation, as well as it is considered as an important factor in the improvement of a society's health, education and environment (Otero & Alejandro, 2016). The Regional Science, Technology and Innovation Systems (RSTIS) manifest particular conditions and challenges of an economic, social, environmental, cultural and political nature, therefore there is a constant need to strengthen and promote the strategic articulation of its actors for providing value to society through the use and management of knowledge (ECoS-CTeI, 2023b). However, the access and use of official bibliographic information on academic and scientific research that is generated in the regions by means of researchers and research groups, has multiple difficulties and limitations as the information is presented in a static, individual and unstructured manner in the official national platforms (Ministerio de Ciencia Tecnología e Innovación, 2023a), (Ministerio de Ciencia Tecnología e Innovación, 2023g). In this order of ideas, the inefficient and restrictive access to official research data in the different territories of Colombia, such as Cauca, hinders their understanding and study. These aspects are vital for strategic decision making of an administrative, academic, social and corporate nature between University, Business, State and Society relationships (Región Administrativa y de Planificación del Pacífico, 2020).

The Ministry of STI (Science, Technology and Innovation) or "Minciencias", through its official tool for the measurement of STeI information called "La Ciencia en Cifras" (Ministerio de Ciencia Tecnología e Innovación, 2023f), does not provide updated and accurate statistics in order to analyze the research environment in detail, since several time periods are omitted and the most recent date available corresponds to the year 2021. Also, the official recording platforms do not allow the efficient exploration or collection of

scientific production data in Cauca and other regions. Currently, these are complicated tasks since "CvLAC" (Ministerio de Ciencia Tecnología e Innovación, 2023c), (Ministerio de Ciencia Tecnología e Innovación, 2021b) and "GrupLAC" (Ministerio de Ciencia Tecnología e Innovación, 2023e), (Ministerio de Ciencia Tecnología e Innovación, 2021c), only allow visualizing bibliographic data through research profiles in single web pages.

The above information evidences the need for continuous effort by the RSTIS actors to strategically identify and consolidate results, competencies and capacities in terms of scientific research, with the aim generating innovation and development in the regions. Therefore, the suitable provision of consistent, accessible and organized information, allows a better analysis of the data to lead a suitable decision making for information management, validation of experiences, interaction between entities and productive development strategies (Calvo Giraldo, 2018), (Tello Clavijo, 2019). From a general framework of the state of the art, several highly recognized tools were identified to support similar tasks (Dattolo & Corbatto, 2022), (Patil et al., 2020), (Ruiz-Rosero et al., 2019), (Aria & Cuccurullo, 2017), (Cobo et al., 2012). Nonetheless, the input stage is done manually by a user from global bibliographic data sources. Although they offer organized data structures for downloading, they include only a small part of the national scientific production, i.e. the few elements that were recognized in high impact journals. Besides, some important entities such as research groups are not recognized in the stored data.

Some lower impact tools, although better related to the addressed context, proposed partial solutions for problems related with platforms such as Lattes from Brazil, which belongs to the same network as CvLAC and GrupLAC (Escarassatti & Bíscaro, 2022), (Mena-Chalco & Junior, 2009) and (Alves et al., 2011). They implemented data extraction functions through text mining, web scraping and visual exploration of multidimensional data from resumes and group profiles in a similar manner as the targeted case for this research. This was done with the purpose of arranging the information to facilitate the analysis. On this same idea, a work done on GrupLAC (Galeano Durán & Prada Pérez, 2019) proposes a user-operated wizard to extract data from the web of a single research group, and then join it into a recommendation system oriented to improve its category by means of certain activities advices.

Based on related works, this research introduces a software for the extraction, structuring and persistence of bibliographic data about the scientific activity registered by researchers and research groups in the CvLAC and GrupLAC platforms in the Colombian territories. The implemented solution seeks to support individuals and entities in the scientific and research environments to be able to collect and manage the bibliographic production data in a simpler and more efficient way, thus contributing to the generation of more and better analyses that are commonly addressed in research activities or decision making within the context of each RSTIS. The software was applied to the Department of Cauca in order to obtain numerous data sets and comparative statistics. It can also be used to generate data sets for any department in the country. The data can be massively updated or extracted, although there are on-demand functions offered to non-technical users. The applied methodology was an adaptation of the Project Management Body of Knowledge (PMBOK) defining the project life cycle, development approach, work breakdown structure and, finally, four phases for the analysis, design, coding and evaluation of the system. This research article presents the work done through different sections explaining the fundamental concepts, the addressed context, hardware and software materials used, methodology, analysis of the results and conclusions.

1. REFERENTES.

1.1. CvLAC

The CvLAC application, which stands for Curriculum Vitae of Latin America and the Caribbean, is an online tool for the registration of resumes of people involved in science,

technology and innovation activities into the "ScienTI" network. These are generally people recognized as researchers or as members of a research group endorsed by the Ministry of Science, Technology and Innovation - Minciencias. Personal and professional information can be registered on the platform, as well as the research products generated, such as articles, books, book chapters, conferences, papers, patents, standards, regulations, courses taught, undergraduate or graduate projects, participation in evaluation committees, software, among others (Ministerio de Ciencia Tecnología e Innovación, 2023d), (Ministerio de Ciencia Tecnología e Innovación, 2021b).

1.2. GrupLAC

The GrupLAC application, acronym for Groups of Latin America and the Caribbean, is a tool made available by the Ministry of Science, Technology and Innovation with the objective of promoting the participation of research groups registered in the "ScienTI" network (Ministerio de Ciencia Tecnología

e Innovación, 2023d). GrupLAC has been created for the integration and exchange of information in the countries of the Latin American and Caribbean territory. Its main objective is to compile and present research groups information based on the records generated by the CVLAC profiles of their active members. In this way, it makes it possible to establish and maintain an updated directory of active groups (Ministerio de Ciencia Tecnología e Innovación, 2021c).

1.3. Web Scraping

The main objective of this practice is to collect data from a web source using an automated program that queries a web server, extracts the necessary data and analyzes it for further use. It provides access to data that can be viewed in a web browser and stored in a database for further processing. It also requires various programming techniques and technologies, such as data analysis and information security, making it a practice that involves extensive technical knowledge. (Mitchell, 2015)

1.4. Regular Expressions

Regex (Regular Expressions) are patterns used for advanced matching of character combinations in text strings. In programming languages such as Python, libraries are used that provide operations suitable for implementation (Python Software Foundation, 2023).

1.5. Regional Innovation Ecosystems

An innovation ecosystem is a dynamic and multifaceted environment where different actors and elements such as companies, institutions, individuals, technologies and culture, interact to drive economic and technological development. Its structure fosters the development by allowing the interaction between suppliers and innovation demanders with a strategic public. In this order of ideas, the innovation ecosystem is configured as a network that includes links to all interested parties, including consumers, service providers, companies, among others. The "Regional Innovation Systems" are environments where regional actors share resources, knowledge and experience to consolidate innovation initiatives through inter-organizational interactions that lead to the production of results. It is relevant to emphasize that educational institutions, and especially universities, play an important role in fostering the creation of innovative networks to accelerate collaborative investments in focused areas and create value (Tello Clavijo, 2019).

1.6. Regional Science, Technology and Innovation System of Cauca

The definition of Regional Science, Technology and Innovation Systems is linked to a geographically determined space and to the notions of the territory, emerging as an evolution and adaptation of the National Science, Technology and Innovation Systems and contemplating a regional innovation ecosystem with cooperation mechanisms such as University, State, Society and Corporation relationships. In this way, the RSTIS are in charge of analyze the different elements that characterize a region in terms of its capacity for development and innovation, given a unique context for each region that determines the dynamics of a specific society (Ministerio de Ciencia Tecnología e Innovación, 2023g). For the case of the Department of Cauca, the set of networks of agents that are articulated in an innovation strategy may involve indigenous people, writers, artists, educators, researchers, ethnic groups, businesses, traders, higher education institutions, peasants, farmers, schools, military forces, the government and the population in general. In this way, an urban, rural, multiethnic and multicultural society is evidenced in a territory of great biodiversity that also has a recognized academic, scientific and technological capacity. Initiatives such as "ECoS-CTeI", bring together actors of the RSTIS of Cauca to contribute value to society from knowledge with the implementation of strategies, instruments, mechanisms and more collaborative projects (Hicks et al., 2015).

1.7. Object-Relational Mapping

It is a programming technique used as a linking mechanism between classes of an object-oriented programming language and the tables of a relational database management system. There are different strategies for mapping associations, inheritance and other relationships between the tables of a relational database. Using this type of database to persist objects that come from object-oriented languages is a common approach and represents a semantic communication mechanism.

2. CONTEXTUAL FRAMEWORK

The articulation between universities and the social, productive and state sectors, has been implemented over the years through actions focused on the construction of regions. In the case of the Department of Cauca, projects and organizations such as "InnovAcción Cauca" have carried out sensitization processes regarding the importance of the relationship and networking of these entities. As a result of these efforts, joint territorial planning exercises have been carried out, such as "Visión Cauca", the Regional Competitiveness Plan, the Departmental Strategic Plan for Science, Technology and Innovation, "ConCiencia Cauca" and "Cauca Emprende", which have contributed to the strengthening of the RSTIS. Several efforts have also been made to strengthen research groups by promoting alliance frameworks between them and entities of the business, social or governmental sector that respond to real needs of the territory (Mosquera Echeverry et al., 2019).

Another project of great importance at the regional level is "ECoS-CTeI", which aims at "participatory action research with the purpose of facilitating joint work among the various RSTIS actors" (ECoS- CTeI, 2023b). This has offered various contributions in the form of tools, knowledge, experiences and training through its strategic lines and its networking components between 2021 and 2023 (ECoS- CTeI, 2023a). Given this scenario, the analysis of information related to scientific production at the regional level by research groups can contribute to the strengthening of the described area.

Researchers' resumes and research group profiles represent the history of scientific, academic and professional activities, and are exhibited as material for analysis on interesting approaches in terms of numbers, distribution, participation, activity, evolution, collaboration, exploration, impact, etc. Finally, it is worth mentioning that this type of information is generated and articulated according to the conditions and needs manifested by each department of the country and symbolizes a similar potential for each

RSTIS.

3. MATERIALS

The articulation between universities and the social, productive and state sectors, has been implemented over the years through actions focused on the construction of regions. In the case of the Department of Cauca, projects and organizations such as "InnovAcción Cauca" have carried out sensitization processes regarding the importance of the relationship and networking of these entities. As a result of these efforts, joint territorial planning exercises have been carried out, such as "Visión Cauca", the Regional Competitiveness Plan, the Departmental Strategic Plan for Science, Technology and Innovation, "ConCiencia Cauca" and "Cauca Emprende", which have contributed to the strengthening of the RSTIS. Several efforts have also been made to strengthen research groups by promoting alliance frameworks between them and entities of the business, social or governmental sector that respond to real needs of the territory (Mosquera Echeverry et al., 2019).

Another project of great importance at the regional level is "ECoS-CTeI", which aims at "participatory action research with the purpose of facilitating joint work among the various RSTIS actors" (ECoS- CTeI, 2023b). This has offered various contributions in the form of tools, knowledge, experiences and training through its strategic lines and its networking components between 2021 and 2023 (ECoS- CTeI, 2023a). Given this scenario, the analysis of information related to scientific production at the regional level by research groups can contribute to the strengthening of the described area.

Researchers' resumes and research group profiles represent the history of scientific, academic and professional activities, and are exhibited as material for analysis on interesting approaches in terms of numbers, distribution, participation, activity, evolution, collaboration, exploration, impact, etc. Finally, it is worth mentioning that this type of information is generated and articulated according to the conditions and needs manifested by each department of the country and symbolizes a similar potential for each RSTIS.

4. MATERIALS

This section presents the tools and resources in terms of hardware and software used in this research work. Table 1 contains the characteristics of the elements selected for the construction of the system. The software consists of the programming language necessary to develop the different stages of the system, the programming frameworks necessary for the development of web applications, the database management systems for persistence, the object-relational mapping tool, the version control tool, the execution environments and the required libraries.

Table 1. Software requirements for the project development

Function	Software	Version
Environments	Python	3.8.16
	Anaconda	2.3.2
	Jupiter Notebook	6.5.2
	Visual Studio Code	1.76.2
Version Control System	Git	2.31.1
	Github	Online
Data Extraction	Beautifulsoup4	4.11.2
	Pyquery	2.0.0
	Request	2.25.1
	Selenium	4.8.2
	Regex	2022.10.31

	Lxml	4.9.1
Data Processing	Numpy	1.24.2
	Pandas	1.5.3
	Openpyxl	3.0.10
	Unittest	3.8.16
Persistence	Pgadmin4	5.5
	PostgreSQL	13.4
	Psycopg2	2.9.3
	SQLAlchemy	2.9.3

Source: own elaboration

5. METHODOLOGY

The applied research methodology was adapted from PMBOK, taking the most current edition of the PMBOK guide as a reference element (Project Management Institute, 2021). This guide delves into the principles of project management and performance domains for delivery effectiveness. In the first instance, an hybrid development approach was defined for the project development, regarding both adaptive and predictive qualities since the existence of a tool capable of handling the raised technical conditions was not identified. Secondly, the project was provided with four development phases for the definition of requirements, modeling, construction and evaluation. Besides, an iterative-incremental strategy was defined within the chosen approach, so that each phase represents an iteration that complements the previous one within each increment until the complete functionality is obtained. The development phases of the modules were defined by:

- Analysis Phase: Allows defining the module requirements and understanding the data or subject matter of analysis.

- Design Phase: It is responsible for the modeling and the functionality representation, based on the requirements.

- Coding Phase: It consists of the programming process of the proposed design and its adaptation to the considered environments.

- Evaluation Phase: It consists of the verification and validation of the requirements, conducting a value judgement.

Ultimately, the project life cycle and the work breakdown structure were constructed from the defined development phases. Thus, each of the phases represents a different work package, each of these in turn containing various sub-packages equivalent to the broken-down activities. The phases and activities developed for the elaboration of this research project are described below.

4.1. Analysis Phase

The initial definition of requirements involved the stakeholders, task analysis, document reading, information synthesis, scope estimation and estimated bibliographic data sources identification. It is worth mentioning that the analysis phase was also nurtured from the progress, findings and particular considerations of the incremental iterations of the methodology. The development of the phase began with the described problematic, the contextual framework of the Cauca RSTIS, and the role of bibliographic data. The gaps in the related works were found from their systematic review and synthesis. The objectives of this research were established based on the construction of a software as a proposal or solution. With this, the knowledge base was consolidated, which allowed to state the appropriate orientation, simultaneously with the applied methodology. Finally, some decisions were made for giving flexibility to the project as a response to the complexity and contingent uncertainty. The first phase of analysis is explained in detail below.

4.1.1. Data Sources Analysis

The defined topic focused on the extraction of data from the sources available in the ScienTI network. It is used on a daily basis by researchers and research groups at national and regional levels through CVLAC and GrupLAC applications. These applications are accessible via Internet and retrieve information from researchers' resumes and research groups' profiles, thus offering a great variety

and quantity of research data from Colombian territories. However, the profiles can only be observed individually and the visualization of the information in both platforms is presented through web pages and table structures in DOMs (Document Object Models) with an HTML (HyperText Markup Language) architecture. The tables are stacked one after the other and are composed of rows and columns. However, their content is unstructured as their various features and values are mixed into text strings. This information corresponds to the data registered by the users and it is detailed in the CVLAC and GrupLAC manuals (Ministerio de Ciencia Tecnología e Innovación, 2021b), (Ministerio de Ciencia Tecnología e Innovación, 2021c) which contain around 80 possible tables for each application.

The volume of the information at a departmental level is estimated from the total number of research groups recognized by Minciencias. A total of 118 groups were identified into the department of Cauca, also including the researchers who were members of these groups. This information was taken from the Minciencias search services (Ministerio de Ciencia Tecnología e Innovación, 2023b). This source is key for retrieving links and lists of groups from other departments in order to extract their bibliographic data using the system proposed in this research.

4.1.2. Data Selection

The available information within the analyzed sources is abundant. For this reason, a selection of tables was carried out based on the "Global Weights of the Products" that are the result of the "research, technological development and innovation processes" consigned in the last "National Call for the recognition and measurement of research, technological development or innovation groups and for the recognition of researchers of the national system of science, technology and innovation - 2021" (Ministerio de Ciencia Tecnología e Innovación, 2021a). The intention was to prioritize the tables containing information of greater weight for the recognition of a research group considering the main product types. Subsequently, the most common interests expressed by some actors of the RSTIS of Cauca and members of the "ECoS-CTeI" project, collected in held meetings, were considered. Also, tables presenting basic information, identifiers and necessary data of the researchers or research groups were included. Tables 2 and 3 presents the selected information for the extraction, structuring and persistence of data from CvLAC and GrupLAC.

Table 2. Selected Tables for the CvLAC Data Extraction

Selected tables	Selected features
Basic Information	CvLAC code, researcher's category, preferred name, citation name, nationality, sex.
Author identifiers	CvLAC code, identifier, link.
Academic education	CvLAC code, type, institution, title obtained, date, name of the project carried out.
Complementary education	CvLAC code, type, institution, title obtained, date.
Activity areas	CvLAC code, activity areas
Languages	CvLAC code, language, speaking level, writing level, reading level, listening level.
Lines of research	CvLAC code, research line, activity status.
Acknowledgments	CvLAC code, title, date.
Jury in evaluation committees	CvLAC code, name, title, type, academic program, orientated name, keywords, knowledge area, involved sectors.
Peer-review	CvLAC code, subject, peer-review type, publisher, journal, institution, application date.
Articles	CvLAC code, authors, name, type, verification status, location, journal, ISSN,publisher, volume, issue, pages, publication date, DOI, keywords, involved sectors.
Books	CvLAC code, authors, book title, type, verification status, location, publication date, publisher, ISBN, volume, pages, keywords, knowledge area, involved sectors.
Book chapters	CvLAC code, authors, chapter, book, location, verified, ISBN, publisher, volume, pages, date, keywords, knowledge area, involved sectors.
Prototypes	CvLAC code, author, name, type, verified, commercial name, contract, registration, location, date, keywords, knowledge area involved, sectors.
Academic social networks	CvLAC code, social network, link
Technology-based companies	CvLAC code, authors, name, type, NIT, chamber of commerce registration, verification status, keywords, areas, sectors.
Softwares	CvLAC code, author, name, type, verified, commercial name, contract or registration number, location, date, platform, environment, keywords, knowledge area, involved sectors.
Business management innovation	CvLAC code, author, name, type, verified, commercial name, contract or registration number, location, date, keywords, knowledge area, involved sectors.
Technology products	CvLAC code, author, name, type, verification status, commercial name, registration contract, location, date, keywords, knowledge area, involved sectors.
Postdoctoral internship	CvLAC code, internship name, involved entity, knowledge area, start date, end date, description.

Source: own elaboration

Table 3. Selected Tables for the GrupLAC Data Extraction

Selected tables	Selected features
Basic information	GrupLAC code, group name, formation date, location, leader, certification, website, email, group classification, areas of activity, related programs, secondary programs.
Institutions	GrupLAC code, institution name, endorsement status.
Lines of research declared by the group	GrupLAC code, registered lines.
Group members	GrupLAC code, CvLAC link for each member, activity status, activity hours, inscription date.
Doctoral academic program	GrupLAC code, program name, date, administrative act number, institution.
Master academic program	GrupLAC code, program name, date, administrative act number, institution.
Other academic program	GrupLAC code, program name, date, administrative act number, institution.
Doctoral courses	GrupLAC code, course name, date, administrative act number, academic program.
Master courses	GrupLAC code, course name, date, administrative act number, academic program.
Published articles	GrupLAC code, verified, type, name, location, journal, ISSN, publication date, volume, issue, pages, DOI, authors.
Published books	GrupLAC code, verification status, type, name, location, publication date, ISBN, publisher, authors.
Published book chapters	GrupLAC code, verification status, book type, chapter name, location, date, book name, ISBN, volume, pages, publisher, authors.
Other published articles	GrupLAC code, verification status, type, name, location, journal, ISSN, publication date, volume, issue, pages, authors.
Other published books	GrupLAC code, verification status, type, name, location, publication date, ISBN, volume, pages, publisher, authors.
Industrial designs	GrupLAC code, verification status, design type, product name, location, elaboration date, availability status, institution, authors.
Business Management innovation	GrupLAC code, verification status, innovation type, product name, location, elaboration date, availability status, institution, authors.
Pilot plants	GrupLAC code, verification status, product type, product name, location, elaboration date, availability status, commercial name, institution, authors.
Other technology products	GrupLAC code, verification status, product type, product name, location, elaboration date, availability status, commercial name, involved institution, authors.
Prototypes	GrupLAC code, verification status, type, name, location, elaboration date, availability status, institution, authors.
Softwares	GrupLAC code, verification status, product type, product name, location, elaboration date, availability status, commercial name, institution, authors.

Technology-based companies	GrupLAC code, verification status, company type, name, registration date, NIT,involved institution, authors.

Source: own elaboration

4.1.3. Data Collection

By going through each research group in their GrupLAC profiles, it was possible to find the members belonging to those groups and then access their CvLAC profiles. The data was collected by iterating the list of groups and the list of members of each group, considering the selected tables and persisting the data in databases. To make this possible, it was necessary to develop functions for the individual extraction of researchers' resumes and groups' profiles, regarding the focused list for Cauca (Ministerio de Ciencia Tecnología e Innovación, 2023b). One of the main points considered in the data collection process was the possibility for updating data and thus strengthen the reliability of the tool and the project. Therefore, massive extraction processes can be accomplished by users with technical knowledge in scripts execution. These processes can take several hours and limits access to users without technical knowledge. For this reason, smaller on-demand updates or extraction tasks can be accomplished by non-technical users through an implemented Graphic User Interface (GUI). It allowed users to extract resumes or research profiles selectively by an easy-to-use web application.

4.1.4. Requirements Definition

The suggested requirements from the analysis phase were as follows:

1. Extract the selected tables for all the researchers and research groups in the Department of Cauca.
2. Extract the selected tables for a CVLAC profile.
3. Extract the selected tables for a GrupLAC profile.
4. Extract the tables selected for the CVLAC profiles of the members of a GrupLAC.
5. Persist the extracted data in the databases.
6. Maintain the level of granularity and the original integrity of the data from their sources.
7. Proper handling of exceptions and notifications during extraction.
8. Assist users with little or no technical knowledge in simple extraction tasks

4.2. Design Phase

4.2.1. Architecture Definition

The defined architecture for the system follows the MVC (Model View Controller) architectural pattern which allows separating the operation logic from the user interface, and facilitates the functionality, maintainability and scalability of the system. Figure 1 shows the defined architecture and technologies.

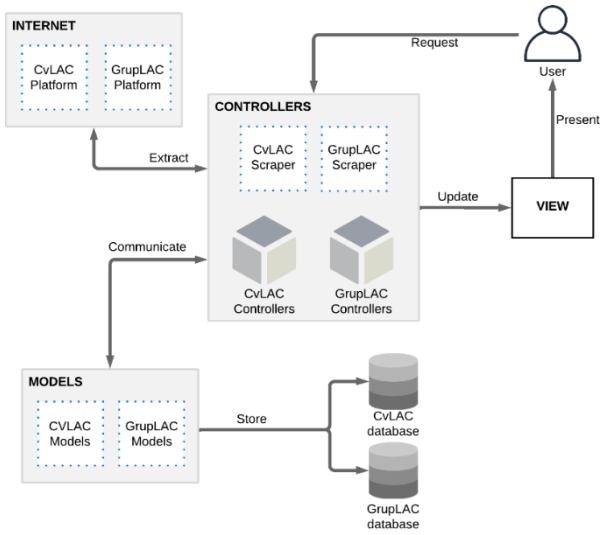

Figure 1. CVLAC-GrupLAC Extractor Module Architecture.
Source: own elaboration

Models oversee the access of the databases' tables and represent the data structures that are stored or consulted. They act as intermediaries between databases and controllers, and each model belong to a unique table or data structure. Two databases were defined, one for CVLAC and the other for GrupLAC. The controllers handle the user interactions, request data from the extractors, process this data and perform the data persistence tasks with the help of the models. Each of the controllers manage one of the databases' tables and interacts specifically with the model that represent the table. There are several extraction methods that retrieve and organize data from the targeted internet platforms following the user's requests. Each web table stored in these platforms is handled by a particular algorithm, so that there is an algorithm for each selected table. The view receives reports from the controllers after fulfilling their tasks and displays them to the user in a human- readable form. Also, the view is designed to be used by a user without technical knowledge. Massive extraction tasks are heavy and time-consuming. They also compromise the integrity of the system and require a certain minimum skill or understanding of the tool.

4.2.2. Programming Model

For data extraction, object-oriented programming model was chosen, also defining classes in charge of the extraction. The attributes of the extracting classes correspond to the information tables of the data sources, so that they contain the unique characteristics or columns of the tables, as well as their records or rows. The methods of the classes contain the unique algorithms to be executed on the tables represented by the attributes. Each driver-model pair of the architecture represents one of these elements for defining and operating the databases' tables. To assist non-technical users, a small web application was developed to provide a user-friendly graphical interface in order to accomplish basic data extraction tasks in a few steps.

4.3. Coding Phase

4.3.1. Data Extraction Stage

During the coding of the data extraction algorithms for the CVLAC and GrupLAC applications, some unique combinations of web scraping techniques, regular expressions and data structures were used in each method using the Python language. Following the proposed

design, the phase started by programming the extractor classes. The data was taken from the DOM documents of each research profile addressed. The algorithms or methods implemented for the tables are similar in their inputs and outputs, although internally they could be considered as black boxes due to their intricate and particular programming. These unique algorithms address particular data engineering challenges to identify, extract and organize the data from internet. In addition, they have the capability for handling possible exceptions or errors.

4.3.2. Data Persistence Stage

The required coding for data persistence consisted of programming the controllers and models using Python, the SQLAlchemy object-relational model tool and the PostgreSQL database management system. The first developed elements were the models that represent the databases' tables of CvLAC and GrupLAC. These models were coded considering the characteristics of the tables, the formats of their values, the relationships between keys, the size of the data and the properties of the fields, among other protocols required by the object-relational strategy. The models are managed by the controllers that receive specific data structures from the extraction methods and operate each specific table of the databases by insert and delete functions. Once the persistence algorithms were programmed, several test extractions were performed so as to correct errors and handle exceptions. Finally, various techniques were implemented to clean and maintain the original integrity of the data taken from the sources.

4.3.3. Graphical User Interface

The developed GUI allowed to extract one resume from a CvLAC link, extract one research group profile from a GrupLAC link, and extract a set of resumes that belong to one research group. Extraction methods performed during a user session, update the databases tables if the identified

codes are already present, otherwise the data is inserted directly. This is a type of on-demand update that consists of the erase of previous data related to a CVLAC or GrupLAC code, and its subsequent extraction and storing. The web application notifies the user through messages about the status of the operations, which may be successful or present some kind of exception related to the Internet connection, the status of the ScienTI platforms, invalidity of the link, etc. The user can display a panel of instructions that works as a support guide. Figure 2 presents the system GUI.

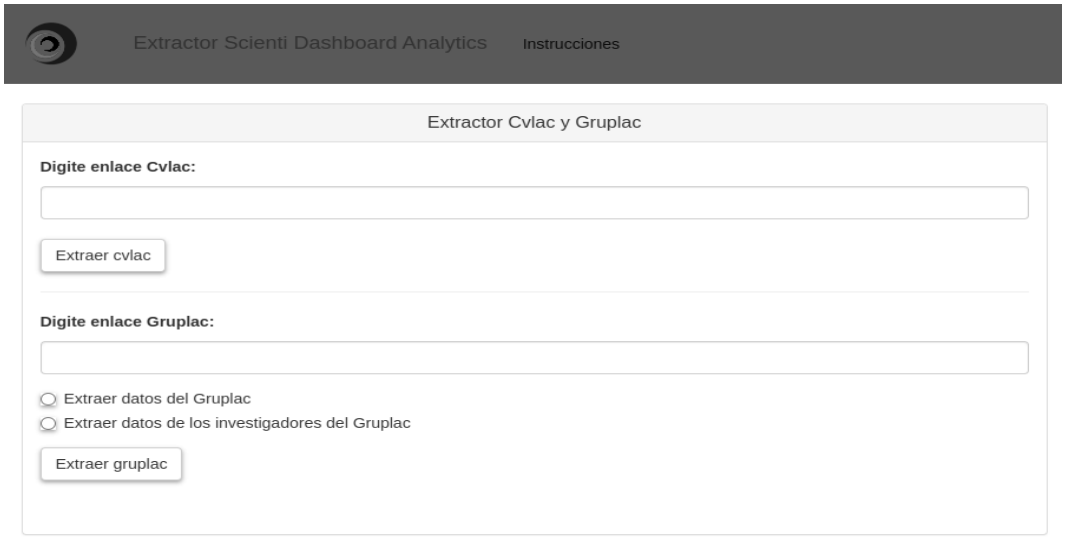

Figure 2. CVLAC-GrupLAC Extractor Module Graphical User Interface.
Source: own elaboration

An administrator user can perform the massive extraction of data and the generation of general statistics directly from the project code by simply executing a script. This process can take several hours depending on the available computing resources and internet connection. The system allows the extension of the proposed scenario since it is possible to extract any CVLAC and GrupLAC at the national level, as well as it can operates links containing a list of GrupLACs from any department of Colombia.

4.4. Evaluation Phase

4.4.1. Verification

To find out whether the subsystem was built correctly, the performance was tested from a technical point of view. This was done by using "black box" unit tests to compare the input and output data of all the algorithms for extracting research profiles. This type of testing was chosen because the developed algorithms presented an intricate composition of logic details, but the characteristics of their inputs and outputs were well known and could be evaluated. Unit tests were executed separately for each method of the coded classes. The verification tests were programmed in order to compare the tables' data structures and yield the percentage of similarity between them based on

the number of records that presented the exact content throughout all their features between pairs of files.

The tables resulting from the extraction methods, which were retrieved and processed from the Internet, were also compared with a set of tables extracted manually from research profiles. Three links were randomly assigned to each selected table from the CvLAC and GrupLAC. For each link, the specific table assigned to that profile was manually constructed by looking at the web page and digitally copying the information from each column and row with the cursor. This process was repeated until covering all the selected tables, resulting in a total of 60 manually extracted files and 60 automatically generated data structures for CvLAC.

The above process was implemented to GrupLAC, resulting in a total of 63 manual generated files and 63 automatic generated data structures. With everything ready, the manually taken elements were organized for being compared with the elements that were automatically generated from the extraction methods. The comparison was done through the 123 unit tests programmed with the Python "unittest" library, 60 tests for CVLAC and 63 for GrupLAC. Each generated table, automatic or manual, contained a variable number of up to 18 columns, depending on its characteristics, and a variable number of rows, depending on its number of records. The results are mentioned in the section of the same name of this article.

4.4.2. Validation

This activity consisted in the validation of the requirements proposed for this system in order to confirm whether it was properly built. Therefore, it was sought to check whether the results matched the analysis phase aims. For this purpose, the assistance of two researchers from the University of Cauca was requested. The participants were considered due to their high academic level in engineering (MSc and PhD) and their high degree of relationship and experience with the raised topic. They carried out the inspection of the tool for a first and a second validation (respectively) of the fulfillment for each requirement. The process consisted of meetings and dialogues held to freely use the tool in order to approve each requirement and provide observations to be taken into account in future work. The results are described in the next section of this article.

5. RESULTS

The developed software can be accessed through the shared repository (Mosquera Perdomo & Salazar Galindez, 2023). The "README" file must be read to install the project following the indicated steps. As expressed in the Coding Phase of the Methodology, Figure 2 presents the graphical interface generated for users with little

or no technical knowledge to use the tool in an on-demand update mode. On the other hand, users with greater technical knowledge can execute massive extractions of information concerning the scientific activity of one or several departments following the instructions of the repository. In this way, numerous complete, organized and updated data sets can be generated for using them in analysis or exploration activities. For this research, the use of the tool was applied to the Department of Cauca.

The system evaluation was successful. The monitoring tool built within the project to verify the correct operation of the system yielded an absolute match for all unit tests executed in both CvLAC and GrupLAC. The result of the verification tasks can be found in the shared repository in the directory indicated by the instructions. It is possible to run these tests freely at any time following the indicated instructions. However, it should be noted that the manually constructed files contain the information as of April 2023 and any future updates on the internet profiles would not be represented in these, but only in the automatically generated files. Consequently, any changes added to the considered profiles will result in differences with respect to the manual files, making it necessary to rework them if future verification tests are meant to be performed.

Regarding the validation process, all the requirements were approved in the two validation stages, which were carried out by two expert engineers from the Faculty of Electronic Engineering and Telecommunications of the University of Cauca. As observations, the possibilities of handling exceptions in the installation processes of the system and avoiding the storage of duplicate values in the databases were expressed. These considerations were accepted after the validation process and were coupled to the source code of the final system.

Finally, general statistics and findings were obtained from the data sets generated for the department of Cauca, and a contrast review was made with respect to the official statistics from "La Ciencia en Cifras" of Minciencias tool (Ministerio de Ciencia Tecnología e Innovación, 2023f), regarding the date of July 10 of 2023. The statistics generated by the built system can be observed through the console of the system after executing the massive extraction and persistence of data from a department. In addition, the addressed datasets were shared in the mentioned repository. The comparative points considered are shown below.

- The total number of researchers identified in Cauca by Minciencias was 370 in 2021. This research identified a total of 3062 researchers for the database generated with GrupLAC, showing a significant difference with respect to the official statistics.

- The data extracted from GrupLAC made it possible to identify that at least 62 out of 118 research groups, have duplicated articles in their own profiles. Even articles verified by Minciencias presented this condition, with at least 56 groups having both verified and duplicated articles in their own GrupLAC profile.

- Following the previous point, at least 873 duplicated articles were identified within the groups' profiles, evidencing a situation of bias in the categorization processes that use "verified" metrics as indicators, since they are susceptible for being artificially inflated. Groups were identified with up to 7 times the same article registered and verified in their own profile, with exact titles and DOIs.

- There are categorical variables in GrupLAC, referring to scientific activity, which provide little or null knowledge of the subject in question. An example of this is the “research lines”, as 531 different lines were identified in Cauca, from which only 10 of them were in fact shared. Most of these research lines do not behave as categorical data.

- The quality of some data records was compromised by an erroneous data entry, referring to text with symbols,

characters, accents, capital letters, spaces, line breaks and other misapplied characteristics. Additionally, there is information that is entered erroneously in cases where text was registered as numerical data, dates or boolean variables, affecting data format. Incomplete and null values were also identified throughout the data sets. The aforementioned information presented inconsistencies from their original sources and were rare exceptional cases.

6. DISCUSSION

- The addressed results allowed to point out several important situations that permeate from the nature of the bibliographic information gathering, to the quality of the analyses that can result from the focused tools. In addition, the studied conditions involve the actions of both individuals and entities belonging to the context environment. It is recommended to encourage good practices on the management of bibliographic information to make it accurate, precise and complete, as well as to strengthen the resources available to capture, analyze and conclude strategic aspects in decision making activities derived from this information. On the other hand, the built system has demonstrated competent capabilities in operation and application to enhance the access and use of research information in Colombian territories, being the first tool of its kind focused on the collection, organization, updating and persistence of this data.

7. CONCLUSIONS

- In this research, a system for the extraction, structuring and persistence of bibliographic data of the scientific activity available in CvLAC and GrupLAC in Colombian territories was implemented. During the research, a case study was carried out on the Department of Cauca with the aim of contributing to the accessibility and exploitation of these resources in order to contribute to the conduction of more and better analyses by the RSTIS actors. The general conclusions of this research are presented below:

- In the state of the art, several proposals for the extraction and manipulation of bibliographic data were observed. However, no suitable tools were identified for collecting, updating, persisting and exploring the targeted data. Consequently, this project gains relevance in the research field in Colombia by proposing a tool that includes stages of the bibliographic data flow from its capture to its preparation for analysis, also noting novel capabilities for its scalability by territories and coupling with analytical tools.

- The quality of the data affects the quality of the analysis that could be carried out from it. So that, the abstention from recording bibliographic data, its careless or incomplete entry and the lack of rigorousness of the official resources for their management in the Colombian context, limit the study of the impact and development of the research ecosystem itself. Although it is possible to apply more data cleaning techniques to handle the aforementioned situations, the bias that could be presented in decision making and strategic planning activities derived from this information depends directly on the quality of the records.

- Access to updated and complete scientific research information in Colombia, considering the official platforms, is tedious and complicated. In that sense, the datasets generated in this research, together with the means used for their creation, encourage the regulated and efficient provision of open and organized data for people with or without technical knowledge within the RSTIS of Cauca for academic or administrative activities.The system

evaluation was successful. On the one hand, expert engineers from the University of Cauca were available for the individual inspection of each requirement of the developed software in two separated validation sessions. On the other hand, exhaustive verification processes for each of the implemented functions were implemented. The results were positive in terms of functionality, usability and satisfaction.

- The statistics and findings considered in the presented results, highlighted the need to improve some aspects related to the management of scientific research information in Colombian territories by official platforms.

REFERENCES

[1]. Alves, A. D., Yanasse, H. H. & Soma, N. Y. (2011). LattesMiner: a multilingual DSL for information extraction from lattes platform. Proceedings of the compilation of the co-located workshops on DSM'11, TMC'11, AGERE! 2011, AOOPES'11, NEAT'11, & VMIL'11, 85–92.

[2]. https://doi.org/10.1145/2095050.2095065

[3]. Aria, M. & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis.

[4]. Journal of Informetrics, 11(4), 959–975. https://doi.org/10.1016/j.joi.2017.08.007

[5]. Calvo Giraldo, O. (2018). La Gestión del Conocimiento en las Organizaciones y las Regiones: Una Revisión de la Literatura. Tendencias, 19(1). https://doi.org/10.22267/rtend.181901.91

[6]. Cobo, M., López-Herrera, A. G., Herrera-Viedma, E. & Herrera, F. (2012). SciMAT: A new science mapping analysis software tool. Journal of the American Society for Information Science and Technology, 63, 1609–1630. https://doi.org/10.1002/asi.22688

[7]. Dattolo, A. & Corbatto, M. (2022). 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. https://doi.org/10.1002/asi.24578

[8]. ECoS-CTeI.(2023a). Boletin ECoS-CTeI Enero 2023.

[9]. https://drive.google.com/file/d/1P9YVempXNl2_w12iL56efx473d8Pn-o1/view ECoS-CTeI. (2023b). ECos-CTeI. https://vri.unicauca.edu.co/ecos-ctei/

[10]. Escarassatti, P. S. & Bíscaro, H. H. (2022). Visual representation of bibliographic production data from Lattes Platform. https://doi.org/10.21203/rs.3.rs-1665862/v1

[11]. Galeano Durán, D. F. & Prada Pérez, L. C. (2019). Diseño de un sistema inteligente para estimación de categorización de grupos de investigación a partir de lineamientos definidos por COLCIENCIAS. [Universidad Distrital Francisco José de Caldas]. http://repository.udistrital.edu.co/handle/11349/22537

[12]. Hicks, D., Wouters, P., Waltman, L., de Rijcke, S. & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. Nature, 520(7548), 429–431. https://doi.org/10.1038/520429a Mena-Chalco, J. P. & Junior, R. M. C. (2009). scriptLattes: an open-source knowledge extraction system from the Lattes platform. Journal of the Brazilian Computer Society, 15(4), 31–39.

[13]. https://doi.org/10.1007/BF03194511

[14]. Ministerio de Ciencia Tecnología e Innovación. (2021a). 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, tecnología e innovación. https://minciencias.gov.co/sites/default/files/upload/convocatoria/anexo_1_-

[15]. _documento_conceptual_2021.pdf

[16]. Ministerio de Ciencia Tecnología e Innovación. (2021b). Manual de usuario CvLAC. https://minciencias.gov.co/sites/default/files/ckeditor_files/D103M06%20Manual%20de%20Usuario%20CVLAC%20V01do.pdf

[17]. Ministerio de Ciencia Tecnología e Innovación. (2021c). Manual de usuario GrupLAC. https://minciencias.gov.co/sites/default/files/ckeditor_files/D103M03%20Manual%20GrupLA C%20V01%20do.pdf

[18]. Ministerio de Ciencia Tecnología e Innovación. (2023a). Acerca de la Red SCienTI : ScienTI. http://www.scienti.net/php/level.php?lang=es&component=19&item=1

[19]. Ministerio de Ciencia Tecnología e Innovación. (2023b). Busqueda de Grupos por Departamento: Cauca. https://scienti.minciencias.gov.co/ciencia-war/busquedaGrupoXDepartamentoGrupo.do?sglPais=COL&sgDepartamento=CA

[20]. Ministerio de Ciencia Tecnología e Innovación. (2023c). CvLAC. https://scienti.minciencias.gov.co/cvlac

[21]. Ministerio de Ciencia Tecnología e Innovación. (2023d). Glosario. https://minciencias.gov.co/sites/default/files/upload/glosario-colciencias.pdf

[22]. Ministerio de Ciencia Tecnología e Innovación. (2023e). GrupLAC. https://scienti.minciencias.gov.co/gruplac/

[23]. Ministerio de Ciencia Tecnología e Innovación. (2023f). La Ciencia en Cifras. https://www.minciencias.gov.co/la-ciencia-en-cifras

[24]. Ministerio de Ciencia Tecnología e Innovación. (2023g). Plataforma SCIENTI - Colombia. https://minciencias.gov.co/scienti

[25]. Mitchell, R. (2015). Web Scraping with Python: Collecting Data from the Modern Web. O'Reilly.

[26]. Mosquera Echeverry, D. M., la Torre Solarte, G., Bastidas Gustín, G., Calvo Giraldo, O. & Sandoval Ruíz, S. M. (2019). Construyendo redes para potenciar la innovaciòn en el Cauca. http://www.unicauca.edu.co/innovaccioncauca/node/4141

[27]. Mosquera Perdomo, E. A. & Salazar Galindez, J. D. (2023). Scienti Extractor. https://github.com/JarbyDaniel/ScientiExtractor

[28]. Otero, M. & Alejandro, R. (2016). La investigación como elemento fundamental para el desarrollo de Latinoamérica. Tendencias y perspectivas. revista del CIECAS-IPN, 39, 35–44. http://hdl.handle.net/10469/10950

[29]. Patil, V. H., Bhavsar, S. A. & Patil, A. H. (2020). An efficient author information retrieval tool for bibliographic record analysis. Journal of Intelligent and Fuzzy Systems, 39(1), 341–353. https://doi.org/10.3233/JIFS-191289

[30]. Project Management Institute. (2021). The standard for project management and a guide to the project management body of knowledge (PMBOK guide) (P. M. Institute, Ed.; Seventh edition). Project Management Institute, Inc.

[31]. Python Software Foundation. (2023). re — Regular expression operations. Python

documentation. https://docs.python.org/3/library/re.html

[32]. Región Administrativa y de Planificación del Pacífico. (2020). Sinergia universidad-empresa-estado- sociedad civil como estrategia de innovación para el desarrollo regional. https://rap- pacifico.gov.co/wp-content/uploads/2020/07/WEBINAR-II.pdf

[33]. Ruiz-Rosero, J., Ramirez-Gonzalez, G. & Viveros-Delgado, J. (2019). Software survey: ScientoPy, a scientometric tool for topics trend analysis in scientific publications. Scientometrics, 121(2), 1165–1188. https://doi.org/10.1007/s11192-019-03213-w

[34]. Tello Clavijo, L. I. (2019). Propuesta para la estructuración de ecosistemas regionales de innovación a partir del rol de instituciones educativas con base en el enfoque de gestión por competencias [Universidad Nacional de Colombia]. https://repositorio.unal.edu.co/bitstream/handle/unal/75992/1110537086.2019.pdf