DASHBOARD PARA EL ANÁLISIS Y VISUALIZACIÓN BIBLIOMÉTRICA DENTRO DEL ÁMBITO DE LOS GRUPOS DE INVESTIGACIÓN EN EL DEPARTAMENTO DEL CAUCA

Trabajo de Grado
Modalidad: Trabajo de Investigación

Edison Alexander Mosquera Perdomo
100618021294
Jarby Daniel Salazar Galíndez
100618010828

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Departamento de Telemática
Línea de Aplicaciones y Servicios sobre Internet
Popayán, agosto de 2023

DASHBOARD PARA EL ANÁLISIS Y VISUALIZACIÓN BIBLIOMÉTRICA DENTRO DEL ÁMBITO DE LOS GRUPOS DE INVESTIGACIÓN EN EL DEPARTAMENTO DEL CAUCA

Trabajo para optar al título de Ingeniero Electrónico y de telecomunicaciones

Trabajo de Grado
Modalidad: Trabajo de Investigación

Edison Alexander Mosquera Perdomo
100618021294
Jarby Daniel Salazar Galíndez
100618010828

Director: PhD. Gustavo Adolfo Ramírez González
Codirector: PhD. Cristhian Nicolás Figueroa Martínez

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Departamento de Telemática
Línea de Aplicaciones y Servicios sobre Internet
Popayán, agosto de 2023

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

1. INTRODUCCIÓN

1.1. PLANTEAMIENTO DEL PROBLEMA

La investigación científica es clave para el desarrollo y crecimiento de un país y sus regiones. Se ha demostrado que esta es generadora de riqueza, conocimiento, bienestar e innovación, además de estar presente en el mejoramiento de la salud, educación y medio ambiente de una sociedad [1]. El departamento del Cauca y su Sistema Regional de Ciencia, Tecnología e Innovación (*SRCTI*), manifiestan condiciones y retos particulares de carácter económico, social, ambiental, cultural y político por los cuales existe la constante necesidad de fortalecer y propiciar la articulación estratégica de sus actores para aportar valor a la sociedad caucana mediante el uso y la gestión del conocimiento [2]. No obstante, el acceso y uso de la información bibliográfica oficial sobre la investigación académica y científica generada en la región por parte de los investigadores y los grupos de investigación, tiene múltiples dificultades y limitaciones al presentarse de manera estática, individual y no estructurada en las plataformas nacionales oficiales [3], [4].

Debe señalarse que existen fuentes de datos bibliográficos globales de excelente calidad como *Web of Science* [5] y *Scopus* [6] que ofrecen además su propias herramientas de análisis y visualización de datos. Sin embargo, estas registran solamente la producción de más alto impacto y no están adaptadas para manejar niveles de granularidad específicos entre entidades y tipos de producto que cobran importancia dentro de algunas actividades analíticas en el alcance regional. Un ejemplo de esto es el análisis de la actividad científica por grupos de investigación y la caracterización de productos como prototipos o software en el “Plan Estratégico Departamental de Ciencia, Tecnología e Innovación del Cauca” [7]. Por lo tanto, la información de estas fuentes no está completamente adecuada al contexto tratado en la región al excluir parte de su producción y actores relevantes.

Por su parte, una estrategia efectiva para estudiar y evaluar la actividad científica en general es la bibliometría [8]. Esta puede describirse como la investigación de la investigación, siendo necesaria la utilización de datos bibliográficos en un contexto estadístico, analítico y visual para la eventual extracción de conocimiento. En este orden de ideas, el acceso manual y restrictivo de los datos oficiales en Colombia y en el Cauca obstaculizan su propia comprensión y estudio. Estos aspectos son vitales para una toma de decisiones estratégica de carácter administrativo,

académico, social y corporativo entre las relaciones Universidad, Empresa, Estado y Sociedad (*UEES*) [9].

El análisis procedente del Ministerio de *CTeI* (Ciencia, Tecnología e Innovación) o "Minciencias", a través de su herramienta oficial para la medición de la información de *CTeI* llamada "La Ciencia en Cifras" [10], muestra que se registraron en el Cauca 117 grupos de investigación con un total de 16.064 productos generados y 370 investigadores activos [11], [12]. Sin embargo, no se cuenta con estadísticas y métricas actualizadas y precisas que permitan conocer cómo ha evolucionado el entorno investigativo a detalle, dado que varios periodos de tiempo son omitidos y la fecha más reciente disponible corresponde al año 2021. Por otro lado, no se identifican los niveles de granularidad necesarios para analizar estadísticamente a un grupo de investigación en específico o a sus relaciones y contrastes con los diversos grupos en su entorno territorial, intelectual o académico. En consecuencia, tampoco se puede descifrar su participación e impacto en este contexto. Cabe mencionar que la exploración detallada de la producción científica oficial en la región es actualmente una tarea complicada e ineficiente puesto que los aplicativos CVLAC [13], [14] y GrupLAC [15], [16], como parte de las plataformas nacionales oficiales, sólo permiten visualizar perfiles de investigación completos y de manera individual.

La información anterior evidencia la necesidad de esfuerzo continuo por parte de los actores del *SRCTI* para identificar y consolidar de manera estratégica y eficiente aquellos resultados, competencias y capacidades en términos de la investigación científica con el fin de generar innovación y desarrollo en la región a través del uso y la gestión del conocimiento. En este orden de ideas, la adecuada disposición de información consistente, accesible y organizada procedente de la investigación, permite un mejor análisis y visualización de los datos para conducir una mejor toma de decisiones en cuanto a la gestión de información, validación de experiencias, interacción entre entidades y estrategias de desarrollo productivo e investigativo [17], [18]. Con todo lo anterior expuesto, el presente trabajo de grado busca dar respuesta a la siguiente pregunta de investigación:

¿Cómo analizar y visualizar la información de la investigación científica en el Departamento del Cauca para conducir una mejor toma de decisiones por parte de los actores del Sistema Regional de Ciencia, Tecnología e Innovación?

1.2. OBJETIVOS

1.2.1. Objetivo general

- Implementar un *Dashboard* basado en la extracción y persistencia de datos de las plataformas CVLAC y GrupLAC, y de la base de datos bibliográfica *Scopus,* para el análisis y visualización bibliométrica dentro del ámbito de los grupos de investigación en el Departamento del Cauca.

1.2.2. Objetivos específicos

- Modelar un conjunto de módulos para la extracción, persistencia, análisis y visualización de datos de las plataformas CVLAC y GrupLAC, y la base de datos bibliográfica *Scopus* a nivel regional.
- Construir los módulos propuestos permitiendo su integración y configuración para el flujo de datos.
- Evaluar los módulos propuestos en un contexto de análisis y visualización bibliométrica referente al ámbito de los grupos de investigación en el Departamento del Cauca.

1.3. CONTRIBUCIONES

- Una herramienta de extracción y persistencia de datos para las plataformas CVLAC y GrupLAC.
- Una herramienta de extracción y persistencia de datos para la base de datos de *Scopus*.
- Un *Dashboard* para la exploración, análisis y visualización bibliométrica de los grupos de investigación científica en el Cauca.
- Tres bases de datos con un total de 45 conjuntos de datos referentes a la información bibliográfica de la actividad científica o investigativa en el Cauca.
- Un artículo de investigación en una revista nacional y un artículo de investigación en una revista internacional (Anexo G).

1.4. ESCENARIO DE MOTIVACIÓN

La articulación entre las universidades y los sectores sociales, productivos y estatales se ha venido implementando a través de los años mediante acciones

enfocadas en la construcción de la región. En el caso del Departamento del Cauca, proyectos y organizaciones como “InnovAcción Cauca” han ejecutado procesos de sensibilización frente a la importancia del relacionamiento y trabajo en red de dichas entidades. Como resultado de los esfuerzos se han realizado ejercicios conjuntos de planeación territorial tales como “Visión Cauca”, el “Plan Regional de Competitividad”, el “Plan Estratégico Departamental de Ciencia, Tecnología e innovación”, “ConCiencia Cauca” y “Cauca Emprende” que han aportado al fortalecimiento del *SRCTI*. También se han hecho esfuerzos para fortalecer a los grupos de investigación a partir de promover marcos de alianza entre estos y entidades del sector empresarial, social o gubernamental que respondan a necesidades reales del territorio. [19]

Otro proyecto de gran auge a nivel regional es “*ECoS-CTeI*”, que apunta a la “investigación-acción participativa que tiene como propósito facilitar el trabajo conjunto entre los diversos actores del *SRCTI*” [2]. Este ha ofrecido diversas contribuciones a modo de herramientas, conocimientos, experiencias y capacitaciones a través de sus líneas estratégicas y sus componentes de redes entre los años 2021 a 2023 [20]. Ante este panorama, el análisis de la información relacionada con la producción científica a nivel regional por parte de los grupos de investigación puede aportar al fortalecimiento del ámbito descrito.

En función de lo planteado, se requiere de un foco adecuado al contexto y de un medio para explorar, analizar y visualizar eficientemente los datos disponibles. Las hojas de vida de los investigadores y los perfiles de los grupos de investigación representan el histórico de las actividades científicas, académicas y profesionales y se exhiben como materia de análisis sobre planteamientos interesantes en términos de cantidades, distribución, participación, actividad, evolución, colaboración, exploración, impacto, etc. En este orden de ideas, la presente investigación aborda la construcción de una herramienta de apoyo para los actores del *SRCTI* que permita extraer conocimiento útil para conducir una mejor toma de decisiones y fortalecer el ecosistema investigativo.

1.5. PARTES DE LA MONOGRAFÍA

La presente monografía se encuentra dividida en los siguientes ocho capítulos que reúnen la investigación realizada:

- ***Capítulo 1: Introducción,*** presenta el planteamiento del problema, objetivos, contribuciones y el escenario de motivación de la investigación.
- ***Capítulo 2: Estado actual del conocimiento,*** presenta los conceptos y tecnologías sobre los que se basa la investigación, junto a las experiencias previas relacionadas con este trabajo de grado.
- ***Capítulo 3: Metodología,*** presenta el proceso de desarrollo de la investigación.
- ***Capítulo 4: Módulo extractor CVLAC - GrupLAC,*** presenta el proceso de modelado, construcción y evaluación de una herramienta extractora de datos *web* para las plataformas CVLAC y GrupLAC.
- ***Capítulo 5: Módulo extractor Scopus,*** presenta el proceso de modelado, construcción y evaluación de una herramienta extractora de datos de *Scopus.*
- ***Capítulo 6: Módulo Dashboard,*** presenta el proceso de modelado, construcción y evaluación de una herramienta de análisis y visualización bibliométrica dentro del ámbito de los grupos de investigación en el Cauca.
- ***Capítulo 7: Prototipo y experimentación,*** presenta la integración, despliegue, evaluación y análisis de resultados de un prototipo orientado a los actores del *SRCTI* para la extracción, persistencia, exploración, análisis y visualización de datos bibliográficos de los grupos de investigación en el Cauca.
- ***Capítulo 8: Discusión,*** presenta la síntesis de los resultados de esta investigación, así como los hallazgos y trabajos a futuro propuestos.

2. ESTADO ACTUAL DEL CONOCIMIENTO

En este capítulo se recopilan los conceptos y tecnologías sobre los cuales se fundamenta esta investigación. Del mismo modo, se describen investigaciones previas realizadas y relacionadas con el desarrollo de herramientas de análisis bibliométrico y cienciométrico, extracción de datos bibliográficos, tableros de visualización, cálculo de indicadores y parámetros, entre otros. Finalmente, se presenta en detalle el *hardware* y *software* utilizado para dejar la posibilidad de replicación del presente estudio.

2.1. CONCEPTOS Y DEFINICIONES FUNDAMENTALES

ScienTI

La plataforma *ScienTI* es una red pública internacional de fuentes de información y conocimiento compuesta por 12 países miembros, entre ellos Colombia, Argentina, Brasil, Chile, Ecuador, Cuba, México, Panamá, Paraguay, Perú, Portugal y Venezuela. Su finalidad es la gestión de la información en el área de la ciencia, la tecnología y la innovación en cada país. Recopila y organiza sistemáticamente información acerca de los investigadores, grupos de investigación, instituciones y proyectos en el ámbito de la investigación, innovación y desarrollo tecnológico. También permite referir estos datos para promover e implementar políticas y estrategias, así como servir de base para la distribución del presupuesto destinado al financiamiento de proyectos investigativos y fomentar programas de cooperación entre los países vinculados. [3]

En Colombia, la investigación se organiza en tres niveles de acuerdo con el ente generador de conocimiento: instituciones, grupos e investigadores. Las instituciones pueden avalar varios grupos de Investigación donde cada grupo está compuesto por un equipo de investigadores. Su tarea consiste en llevar a cabo actividades de investigación con el objetivo de generar productos que se evalúan en la plataforma *ScienTI* según el modelo de medición de grupos de investigación de Minciencias. El puntaje obtenido por el producto individual se utiliza para calcular el puntaje total del grupo de Investigación asociado. Este puntaje es fundamental para el grupo ya que determina su clasificación en una escala que va desde "A1" hasta "C", lo que a su vez aumenta la posibilidad de participar en diferentes convocatorias, eventos

científicos, alianzas, además de diferentes beneficios para la institución que respalda al grupo. [21]

La red *ScienTI* en Colombia provee dos sistemas encargados del registro, actualización, procesamiento y presentación de datos con respecto a la información relacionada a grupos y personas que forman parte de las actividades de investigación científica. Estos son llamados *CVLAC* y *GrupLAC.*

CVLAC

La aplicación *CVLAC,* siglas de *Curriculum Vitae* de Latinoamérica y el Caribe, es una herramienta en línea destinada al registro de hojas de vida de las personas que participan en actividades de ciencia, tecnología e innovación inscritas en la red *ScienTI.* Son generalmente persona reconocidas como investigadores o como integrantes de un grupo de investigación avalado por el Ministerio de Ciencia, Tecnología e Innovación - Minciencias. En la plataforma se puede registrar información personal y profesional, así como los productos de investigación generados tales como artículos, libros, capítulos de libros, conferencias, ponencias, patentes, normas, regulaciones, cursos impartidos, tesis de pregrado o posgrado, participación en comités de evaluación, software, entre otros. [22] [14]

El modelo de medición de grupos de investigación de Minciencias define todos los tipos de productos de investigación aceptados por esta entidad, y su clasificación se basa en su naturaleza y grado de impacto [14]. Estos son de gran importancia para los investigadores dado que permiten entender los criterios de evaluación que utiliza Minciencias y orientar sus trabajos de investigación para aumentar la calidad y el impacto de sus productos. En la Tabla 2.1, se listan los elementos que están disponibles al agregar información en la herramienta *CVLAC*.

Elementos	Descripción
Fechas	Permite la introducción de fechas importantes relacionadas con la carrera profesional y académica del investigador, tales como fechas de graduación, obtención de títulos, publicación de trabajos, entre otros.
Ciudad / Municipio	Permite indicar la ciudad o municipio donde se ubica el centro de investigación, institución académica, empresa, organización o grupo de investigación con el que está vinculado el investigador, así como el lugar donde desarrollo el producto.

Palabras clave	Términos o frases que se utilizan para describir los temas de investigación, intereses y especialidades del investigador, y que facilitan la búsqueda y la selección de información de interés.
Áreas de conocimiento	Permite la selección y la asignación de una o varias áreas de conocimiento en las que el investigador tiene experiencia y ha desarrollado trabajos de investigación.
Vinculación de instituciones	Indica la institución, universidad, empresa u organización con la que el investigador está vinculado, y en la que ha realizado actividades de investigación o de docencia.
Vinculación de personas	Permite la vinculación de otras personas, ya sean colaboradores, coautores o estudiantes, que han trabajado con el investigador en proyectos de investigación o en otras actividades académicas.
Programas académicos	Incluye los programas académicos en los que el investigador ha participado como docente, investigador o estudiante.
Reconocimientos	Registra los reconocimientos, premios o distinciones obtenidos por el investigador en su carrera profesional o en su trabajo de investigación.
Coautores	Permite mencionar los nombres de los coautores de los trabajos de investigación desarrollados por el investigador.
Comunidades	Posibilita la pertenencia a comunidades académicas o científicas en las que el investigador participa y comparte conocimientos e información.
Referencia en revistas	Permite incluir referencias de los artículos científicos publicados por el investigador en revistas científicas indexadas.
Referencia en libros	Permite incluir referencias de los libros, capítulos de libros o editoriales publicados por el investigador.

Tabla 2.1. Elementos disponibles en *CVLAC*. Adaptado de [14]

GrupLAC

El aplicativo *GrupLAC,* siglas de Grupos de Latinoamérica y el Caribe, es una herramienta que pone a disposición el Ministerio de Ciencia, Tecnología e Innovación con el objetivo de promover la participación de los grupos de investigación registrados en la red *ScienTI.* Según los lineamientos de Minciencias, todo grupo de investigación debe tener un líder que puede ser cualquier miembro de la institución asociada, siendo este el responsable de dirigir y coordinar las actividades de investigación. Adicionalmente, el grupo debe contar con al menos dos integrantes para su conformación. [22]

GrupLAC al igual que *CVLAC,* ha sido creada para la integración e intercambio de información en los países de la región de América Latina y del Caribe. Su objetivo principal es recopilar y presentar la información de los grupos de investigación a

partir de los registros generados en los perfiles de *CVLAC* de los integrantes. De esta manera, permite establecer y mantener un directorio actualizado de los grupos que se encuentran activos. De acuerdo con la convocatoria nacional para la clasificación de grupos de investigación, se consideran cuatro grandes categorías de productos que estos grupos pueden generar, consignados en la Tabla 2.2. [16]

Productos	Descripción
Productos de Generación de Nuevo Conocimiento	Aportes significativos en un campo de conocimiento que han sido validados y discutidos por la comunidad científica, los cuales han sido incorporadas en la discusión académica y en la investigación y el desarrollo tecnológico. Estos avances pueden ser la fuente de innovaciones y tienen un impacto muy significativo en la forma en que se aborda el tema en cuestión.
Productos Resultados de Actividades de Desarrollo Tecnológico e Innovación	Son el resultado de la generación de ideas, métodos y herramientas cuya importancia radica en su capacidad para impulsar el cambio, mediante la creación de soluciones innovadoras que aborden problemas y desafíos relevantes en una determinada área, fomentando el progreso y el crecimiento en distintos ámbitos, ya que genera impacto positivo en la sociedad.
Productos de Apropiación Social y Circulación del Conocimiento	Son el resultado de una práctica colaborativa en la que actores, ya sea a título individual o conjunta, intercambian saberes y experiencias. Esta dinámica social da lugar a una circulación de conocimiento en la que se promueve la discusión, la evaluación, la aplicación y la transferencia del conocimiento en el ámbito cotidiano, facilitando la innovación y la generación de productos que pueden ser útiles para el bien común.
Productos de Formación de Recursos Humanos	Incluyen, en primer lugar, la creación de espacios donde se brinda asesoramiento y guía a estudiantes durante la realización de sus tesis o trabajos de grado, lo que les permite obtener títulos de doctorado, maestría o profesional. En segundo lugar, se llevan a cabo proyectos de Investigación, Desarrollo e Innovación (ID+I) que incluyen formación y apoyo a programas de formación en distintas áreas de conocimiento.

Tabla 2.2. Tipos de productos en *GrupLAC*. Adaptado de [16]

Lattes

De acuerdo con [23], la plataforma *Lattes* es un sistema de información creado por el Consejo Nacional de Desarrollo Científico y Tecnológico de Brasil, que integra bases de datos de currículos, grupos de investigación e instituciones en un único sistema, por medio del sistema *CVLAC* de *ScienTI*. Su uso se ha extendido a otras agencias de financiación, fundaciones estatales y entidades gubernamentales relacionadas con la ciencia, la tecnología y la innovación. La disponibilidad en línea de estos datos desde el repositorio "*LattesData",* es esencial para la práctica de la

ciencia abierta en Brasil. *Lattes* proporciona una plataforma para la extracción de datos llamada "*Extração Lattes"* accesible mediante permisos previamente solicitados para obtener los datos necesarios en proyectos específicos de investigación. [24]

Scopus

Scopus es una base de datos de información bibliográfica y documentación seleccionada, organizada y presentada por expertos independientes en diversos campos del conocimiento quienes son reconocidos como lideres en sus ámbitos. Pertenece a la firma "*Elsevier"* y cuenta con más de 1.8 billones de referencias citadas a fechas desde 1970 hasta la actualidad, más de 84 millones de registros, más de 17.6 millones de perfiles de autores y más de 94.800 perfiles de afiliaciones o instituciones. Adicionalmente, presenta sofisticadas herramientas analíticas que generan resultados precisos de citaciones, perfiles detallados de investigadores, tendencias de investigación, rastreo de nuevas publicaciones y diversas métricas o indicadores relevantes que conduzcan, dado el contexto, a una mejor toma de decisiones y acciones para un usuario. [6]

La rica arquitectura de metadatos sobre la cual *Scopus* está construida y sus diversas herramientas promueven el flujo de ideas y el reconocimiento de personas e instituciones a través de las publicaciones y su contenido de valor como, por ejemplo, la generación de nuevo conocimiento. También propicia un entorno de redes y conexiones entre sus actores y productos. Entre sus registros se encuentra un mayor volumen de publicaciones en el área de las ciencias sociales con alrededor de un 35%, seguido del área de las ciencias físicas con 27%, luego el área de las ciencias de la salud con 23% y por último el área de las ciencias de la vida con un 15%. [25]

Scopus APIs

El término *API* es una abreviatura de "*Application Programming Interface"* o en español "Interfaz de Programación de Aplicaciones", cuya función principal es comunicar dos componentes *software* a través de una variedad de mecanismos y definiciones programadas para tareas y manejo de información específica. El contenido de *Scopus* captura publicaciones en revistas científicas o editoriales que cuentan con algún grado de reconocimiento aceptado a nivel mundial, así como el

perfilamiento de autores e instituciones. Esta base de datos cuenta con un conjunto de *APIs* que trabajan sobre su voluminoso y variado contenido, y ofrecen las mismas características observables en las interfaces gráficas de su plataforma. No obstante, se presentan en un formato entendible por maquinas que permite ceder a un *software* las tareas de buscar y retornar información de publicaciones, investigadores, instituciones, etc.

El uso de las *APIs* de *Scopus,* como se observa en la Figura 2.1, permite extraer información de manera automática desde su base de datos para luego ser usada por el sistema del usuario. La información se procesa mediante solicitudes que definen parámetros y valores. Cabe mencionar que el uso de dichas *APIs* está restringido no sólo a suscriptores de la firma, si no a proyectos aprobados por un equipo de contacto de la empresa que se asigna a través de la solicitud de un sujeto autorizado por la institución suscriptora. [26], [27].

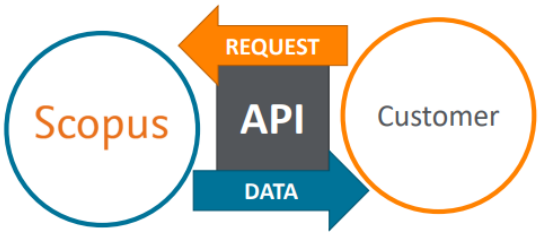


Figura 2.1. Modelo simple del funcionamiento de una *API* de *Scopus*. Tomado de [26]

Las *APIs* de *Scopus* pueden ser usadas en multitud de casos de uso como, por ejemplo, estrategias de investigación, desarrollo, inversión, reconocimiento o evaluación de interés para instituciones e investigadores, además de poder integrarse a sistemas de uso propio. En total existen 12 *APIs* que operan sobre *Scopus* [28] de las cuales fueron necesarias 5 para el desarrollo de la presente investigación. A continuación, se presenta la Tabla 2.3 con la descripción de cada una de las *APIs* usadas y la cantidad de solicitudes semanales permitidas.

API	**Descripción**	**Cuota semanal [JSON *request*]**
Affiliation Search API	Representa las interfaces de búsqueda relacionadas con perfiles de afiliaciones o instituciones. Esta búsqueda se ejecuta sobre un *cluster* de afiliaciones que contiene perfiles de instituciones u organizaciones afiliadas a *Scopus*.	5.000

Author Search API	Representa las interfaces de búsqueda asociadas con la base de perfiles de autor de *Scopus*. Los resultados obtenidos relacionan los perfiles de autores que coinciden con la solicitud a sus afiliaciones actuales.	5.000
Scopus Search API	Representa las interfaces de búsqueda relacionadas con parámetros de *Scopus* referentes a documentos como título, tema, país, etc. Los resultados retornados sugieren coincidencias según las especificaciones de la búsqueda e incluyen algunos metadatos y correspondencias de los productos.	20.000
Abstract Retrieval API	Representa una interfaz de recuperación de datos para el resumen de un documento de *Scopus* y sus metadatos, los cuales incluyen enlaces al autor y afiliación de correspondencia.	10.000
Author Retrieval API	Representa una interfaz de recuperación de datos para el perfil de un autor en *Scopus* el cual incluye indexaciones a publicaciones realizadas por este y perfiles de afiliación relacionados.	5.000

Tabla 2.3. *APIs* de *Scopus* utilizadas. Adaptado de [29]–[34]

Bibliometría

La bibliometría, anteriormente conocida como “bibliografía estadística”, es una herramienta interdisciplinaria que permite realizar un análisis cuantitativo de la producción, difusión y uso de la literatura definida como un conjunto de documentos relacionados por su temática o autoría [35]. La bibliometría utiliza técnicas matemáticas y estadísticas para analizar las publicaciones científicas y académicas. Es útil para estudiar patrones de comportamiento en la comunicación y el uso de la información [36]. El análisis bibliométrico puede incluir diferentes tipos de documentos, como artículos de revistas científicas, tesis de grado, informes técnicos, libros, entre otros. Por otro lado, la bibliometría también puede analizar otros aspectos de la comunicación científica, como las revistas electrónicas, conferencias, la producción de patentes, la colaboración entre autores, entre otros.[37]

El objetivo de la bibliometría es generar conocimiento sobre las tendencias en la producción y uso de la información científica y académica. Esto puede ser de gran ayuda para la toma de decisiones en diferentes áreas como la política, la gestión de

la información, la evaluación de la calidad de la investigación, entre otras. Sobre el mismo marco planteado, la bibliometría se ha utilizado para evaluar la productividad e impacto de los investigadores y las instituciones, lo que ha generado cierta controversia en el ámbito académico. En este sentido, es importante señalar que la bibliometría no es la única herramienta útil para evaluar la calidad de la investigación puesto que existen elementos considerables tales como la creatividad y la relevancia de los resultados obtenidos. Para ello se implementan alternativas como el proceso de revisión por pares, en el cual expertos en el campo temático de la investigación evalúan el trabajo presentado y emiten comentarios, sugerencias y recomendaciones. [8]

Cienciometría

La cienciometría es una disciplina que aplica técnicas bibliométricas a la ciencia con el objetivo de estudiar matemática y estadísticamente los patrones de investigación. Esta disciplina también se conoce como la ciencia de la ciencia, ya que su metodología se basa en el uso de indicadores cuantitativos para caracterizar el comportamiento y la orientación de la investigación científica. A diferencia de las técnicas bibliométricas habituales, la cienciometría puede ir más allá y examinar el desarrollo e incluso la política de las ciencias, permitiendo comparar las políticas de investigación científica de un país a otro, así como la cantidad de dinero o el número de científicos en cada país. Sin embargo, es importante destacar que la cienciometría suele centrarse en las ciencias físicas y naturales, así como en las matemáticas, excluyendo en gran medida las ciencias sociales. [38]

Web Scraping

Aunque el "*web scraping" o "*raspado web*"* no es una técnica novedosa, ha evolucionado a lo largo de los años, cambiando su nombre desde términos como "*screen scraping"*, "*data mining"* o "*web harvesting"* a "*web scraping"*. El objetivo principal de esta práctica es recolectar datos de una fuente *web* utilizando un programa automatizado que consulta un servidor *web*, extrae los datos necesarios y los analiza para su posterior uso. Aunque algunas entidades ponen a disposición *APIs* que pueden proporcionar una forma cómoda y bien estructurada de acceder a sus datos, a veces no son suficientes para todas necesidades, ya sea por limitaciones en el volumen de solicitudes o en el formato de los datos. Es aquí donde el *web scraping* se convierte en una herramienta valiosa, dado que permite acceder

a los datos que se pueden visualizar en un navegador *web* y almacenarlos en una base de datos para su posterior procesamiento. También requiere de diversas técnicas y tecnologías de programación, como el análisis de datos y la seguridad de la información, por lo que es una práctica que implica un amplio conocimiento técnico. [39]

Dashboard

Los *dashboards* fueron introducidos, en sus comienzos, a la literatura del área de estudio del *marketing* bajo el nombre de "*marketing dashboard*" cuya función principal es el uso de los datos de determinadas firmas o compañías con el objetivo de observar métricas clave para sus negocios a través de un tablero digital de forma visual, sencilla y atractiva. Muchas empresas usan los *dashboards* para monitorear su efectividad y guiar la toma de decisiones en sus procesos de negocios en plena era de la información y ante el vertiginoso crecimiento en complejidad y diversidad de los datos del mercado. Estas herramientas rastrean diferentes indicadores de desempeño con relación a sus valores de entrada para ser usados como materia de evaluación y planeación [40]. La Figura 2.2 muestra un ejemplo aplicado al estudio de la pandemia del COVID-19.

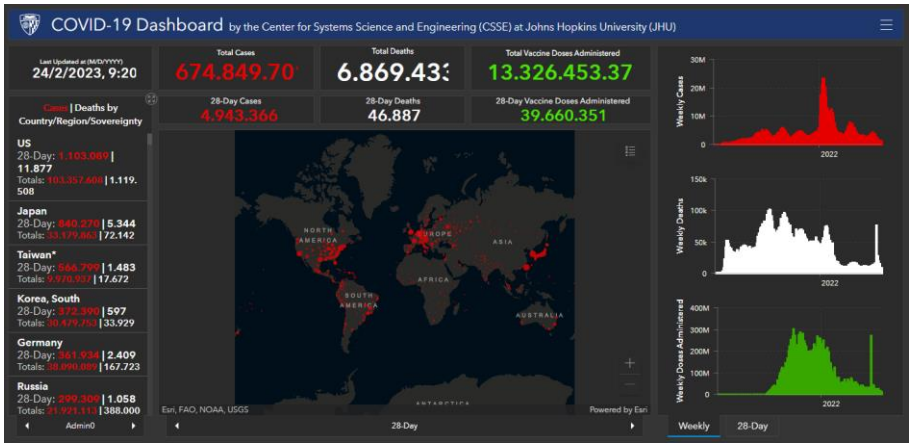


Figura 2.2. Ejemplo de un *dashboard* para el monitoreo del COVID-19. Tomado de [41]

Siguiendo a [40], un *dashboard* se entiende entonces como una herramienta para el monitoreo, gestión, análisis y visualización de datos e indicadores clave para conocer el estado en el que se encuentra una organización o proceso de cualquier naturaleza si se cuenta con los datos suficientes. En este se recopilan variedad de figuras, gráficos, números y textos particularmente distribuidos y organizados con el

fin de transmitir información de valor que a su vez sea digerible por el usuario. También cuentan con funcionalidades de utilidad que permiten manipular y filtrar los datos disponibles.

Ecosistemas de innovación regional

Un ecosistema de innovación es un entorno dinámico y multifacético donde distintos actores y elementos como empresas, instituciones, individuos, tecnologías y cultura, interactúan para impulsar el desarrollo económico y tecnológico. Su estructura fomenta el desarrollo al permitir la interacción entre proveedores y demandantes de innovación con un público estratégico. En este orden de ideas, el ecosistema de innovación se configura como una red que incluye enlaces a todas las partes interesadas, incluyendo consumidores, proveedores de servicios, empresas, entre otros. [18]

Siguiendo a [18], los “Sistemas Regionales de Innovación” son entornos donde los actores del ámbito regional comparten recursos, conocimientos y experiencia para consolidar iniciativas de innovación a través de interacciones interorganizacionales que conducen a la producción de resultados. Es relevante recalcar que las instituciones educativas, y en especial las universidades, desempeñan un papel importante al fomentar la creación de redes innovadoras para acelerar las inversiones colaborativas en áreas enfocadas y crear valor. En la Figura 2.4 se observan los componentes principales de un Sistema Regional de Innovación.

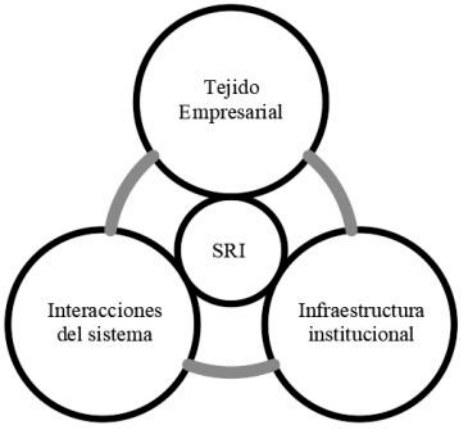


Figura 2.3. Componentes del Sistema Regional de Innovación. Tomada de [18]

SRCTI (Sistema Regional de Ciencia, tecnología e innovación del Cauca)

La definición de los sistemas regionales de ciencia, tecnología e innovación está vinculada a un espacio geográficamente determinado y a las nociones del territorio, surgiendo como una evolución y adaptación de los sistemas nacionales de ciencia, tecnología e innovación y contemplando a su vez un ecosistema de innovación regional con mecanismos de cooperación como las relaciones *UEES.* De esta manera los *SRCTI* son una forma de analizar los distintos elementos que caracterizan una región en cuanto a su capacidad de desarrollo e innovación dado un contexto único para cada región determinante para las dinámicas de una sociedad especifica. [4]

Para el caso del Departamento del Cauca, el conjunto de redes de agentes que se articulan en una estrategia de innovación puede involucrar a indígenas, escritores, artistas, educadores, investigadores, etnias, empresas, comerciantes, instituciones de educación superior, campesinos, agricultores, escuelas, fuerzas militares, al gobierno y a la población en general. De este modo, se evidencia una sociedad urbana, rural, pluriétnica y multicultural en un territorio de gran biodiversidad que además cuenta con capacidad académica, científica y tecnológica reconocida. Iniciativas como *ECoS-CTeI,* reúnen actores del *SRCTI* del Cauca para aportar valor a la sociedad a partir del conocimiento con la ejecución de estrategias, instrumentos, mecanismos y de más proyectos colaborativos [8]. En la Figura 2.5 se observan algunos de los actores del *SRCTI* del Cauca.

Figura 2.4. Actores reconocidos del *SRCTI* del Cauca. Tomado de [20].

DBMS, Database Management System

Database Management System o en español, sistema gestor de bases de datos, es una colección de datos interrelacionados entre sí, acompañada de un conjunto de programas usados para acceder a dichos datos. El papel que juega un *DBMS* es el de proporcionar el almacenamiento y recuperación de la información de las bases de datos de forma oportuna y eficiente, estos sistemas de bases de datos son diseñados para manejar grandes volúmenes de información de diversos formatos y tamaños, manteniendo su integridad y aplicando cierto grado de seguridad. [42]

ORM, Object-Relational Mapping

Object-Relational Mapping o en español, mapeo objeto-relacional, es una técnica de programación usada como mecanismo de enlace entre clases de un lenguaje de programación orientado a objetos y las tablas de un sistema gestor de bases de datos de tipo relacional. Existen diferentes estrategias para mapear asociaciones, herencias y de más relaciones entre las tablas de una base de datos relacional. Usar este tipo de bases de datos para persistir objetos que provienen de lenguajes orientados a objetos es un enfoque común y representa un mecanismo de comunicación semántica. Este paradigma también presenta una variedad de retos superados y por superar desde un punto de vista funcional entre los diferentes sistemas gestores de bases de datos y los lenguajes de programación orientados a objetos. Por ese motivo, han aparecido diversos *ORM frameworks* como *Hibernate, Doctrine* y *SQLAlchemy* enfocados a diferentes lenguajes de programación. [43]

2.2. REVISIÓN SISTEMÁTICA DE LA LITERATURA (RSL)

En la presente investigación se estudió el estado actual del conocimiento acerca del análisis y visualización bibliométrica de la producción científica a nivel nacional e internacional. Se hizo énfasis sobre los grupos de investigación reconocidos a nivel institucional y las plataformas que condensan sus datos de interés. Lo anterior se logró siguiendo la guía establecida por Kitchenham y Charters [44] para la construcción de una revisión sistemática de la literatura apropiada para investigaciones relacionadas con *Software* e ingeniería. Con la revisión sistemática se buscó obtener información de las técnicas y tecnologías empleadas para encontrar las brechas presentes en este campo de investigación siguiendo las etapas de la Figura 2.5.

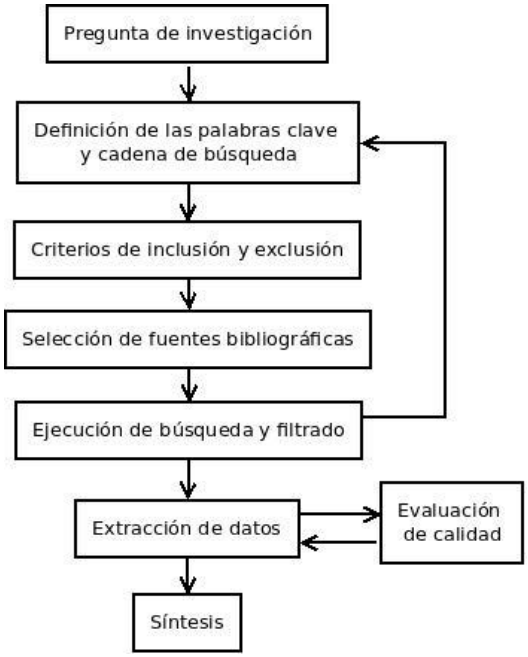


Figura 2.5. Proceso de revisión sistemática de literatura. Adaptado de [44]

Inicialmente, se realizó el planteamiento de la pregunta de investigación de la RSL con base en la pregunta de la sección 1.1 y el escenario de motivación de la sección 1.4: ***¿Cómo realizar el análisis y visualización bibliométrica de los grupos de investigación científica en el Departamento del Cauca para conducir una mejor toma de decisiones por parte de los actores del SRCTI?***

2.2.1. Definición de las palabras clave y cadena de búsqueda

La definición de las palabras clave y las cadenas de búsqueda se realizó en relación con el problema definido y la pregunta de investigación, tomándose conceptos clave y omitiendo tecnologías redundantes. Se decidió construir dos cadenas de búsqueda con el fin de identificar, en primer lugar, la documentación primaria de la materia en cuestión al usar los términos base de: “bibliometría”, “mapeo científico”, “cienciometría” y “bibliografía”. En segundo lugar, se realizó un sondeo a nivel regional y nacional utilizando términos en español y en ingles con el objetivo de observar las investigaciones realizadas sobre la población y plataformas comprendidas en *ScienTI.* Dado que durante algunas iteraciones en la definición de las palabras clave se encontró que no hay trabajos de mayor relevancia acerca de

CVLAC y *GrupLAC* publicados en revistas de alto impacto, se optó por hacer búsquedas en distintas bases de datos esperando obtener aún más resultados relevantes en revistas de menor impacto, o bien, en repositorios universitarios. La primera cadena de búsqueda definida, orientada a la literatura más reconocida fue la siguiente:

Article Title: *(bibliometr* OR "science mapping" OR scientometr* OR bibliograph*) AND (analy* OR extrac* OR visua*) AND (tool OR software OR dashboard)*

AND

Author Keywords: *(bibliometr* OR scientometr* OR bibliograph*)*

La anterior cadena busca coincidencias entre el titulo y las palabras clave. Además, se adhiere el filtro "*Author Keywords"* que es definido por los autores de una publicación refiriendo la temática intencionada de la misma. La segunda cadena de búsqueda enfocada en la literatura menos reconocida, pero con un mayor grado de relación, fue la siguiente:

Article Title: *(cvlac OR gruplac OR colciencias OR lattes) AND (sistema OR plataforma OR herramienta OR análisis OR visualización OR extracción OR system OR platform OR tool OR analysis OR visualization OR extraction)*

La cadena incluyó términos en inglés y en español debido al tipo de resultado que se esperaban obtener al incluir revistas contempladas en Latinoamérica y Caribe.

2.2.2. Criterios de inclusión y exclusión

Se estableció un conjunto de criterios de inclusión y exclusión con el fin de determinar la calidad de los estudios, según se indica en [44]. Se planteó que entre los estudios se deberían excluir aquellos que realizan actividades de análisis bibliométrico sobre variadas temáticas que no corresponden a la bibliometría en sí, y qué hacen uso de herramientas bibliométricas, pero no presentan el desarrollo de dichas herramientas. En la Tabla 2.4 se muestra un resumen de los criterios utilizados.

Criterios de inclusión	• Artículos publicados en revistas científicas. • Artículos que aborden al menos un módulo de extracción, análisis o visualización de datos bibliográficos de la producción científica publicada en revistas de alto impacto. • Artículos o monografías que aborden aplicativos para la manipulación y análisis de datos bibliográficos de las plataformas de *ScienTI*.
Criterios de exclusión	• Literatura gris. • Artículos que presentan un análisis bibliométrico sin enfocar la bibliometría como tema. • Artículos que son sólo capítulos de libros o fueron presentados en conferencias. • Informes breves. • Fuentes secundarias.

Tabla 2.4. Criterios de inclusión y exclusión. Fuente propia.

2.2.3. Selección de fuentes bibliográficas

Se definieron 3 fuentes bibliográficas presentadas en la Tabla 2.5, teniendo en cuenta las observaciones hechas en la sección 2.2.1 de la monografía. Se tuvieron en cuenta algunas de las fuentes bibliográficas con mayor prestigio y más utilizadas para encontrar estudios del ámbito de la tecnología y sus campos de aplicación.

Fuente bibliográfica	URL
Web of Science	https://www.webofscience.com
Scopus	https://www.scopus.com
Google Scholar	https://scholar.google.com

Tabla 2.5. Fuentes bibliográficas utilizadas. Fuente propia.

2.2.4. Ejecución de búsqueda y filtrado

La primera cadena de búsqueda definida en la sección 2.2.1 se aplicó sobre las bases de datos de *Web of Science* y *Scopus*, dado que son fuentes bibliográficas de gran reconocimiento a nivel global e indexan revistas científicas de alto impacto que además cuentan con altos estándares de calidad para seleccionar sus publicaciones. La segunda cadena de búsqueda se ejecutó sobre la base de datos de *Google Scholar*, que permitió obtener resultados de una mayor variedad de revistas y repositorios incluyendo estudios escritos en español y portugués

enfocados a contextos más cercanos al del territorio y a la población de interés. La segunda cadena de búsqueda se adaptó a *Google Scholar* de la siguiente manera:

allintitle: *"cvlac|gruplac|colciencias|lattes" + "sistema|plataforma|herramienta|system|platform|tool|extracción|análisis|visualización| extraction|analysis|visualization"*

Los resultados de las búsquedas realizadas, en términos de cantidades, se listan en la Tabla 2.6.

Fuente bibliográfica	Número de artículos
Web of Science	32
Scopus	77
Google Scholar	389
Total	**498**

Tabla 2.6. Resumen del número de artículos encontrados por fuentes bibliográficas. Fuente propia.

En primera instancia, se identificó que 142 de los 389 resultados de *Google Scholar* eran solamente etiquetas de cita sin enlace a un documento, artículo o producto de investigación, por lo cual fueron descartados como materia de documentación obteniendo así un remanente parcial de 247 documentos provenientes de esta base de datos y un total general de 356 documentos para las tres bases de datos. Posteriormente, se realizó un filtrado de los documentos obtenidos a través de una lectura rápida del título, resumen, palabras clave y contenido en general de los 356 documentos. Adicionalmente, se aplicaron los criterios de inclusión y exclusión definidos anteriormente y se encontraron la cantidad de artículos duplicados, rechazados y aceptados como se observa en la Tabla 2.7.

Clasificación	Número de artículos
Duplicados	51
Rechazados	434
Aceptados	13

Tabla 2.7. Resumen de filtrado de la documentación obtenida. Fuente propia.

Se encontró que la búsqueda entre las bases de datos *Web of Science* y *Scopus* presentó un total de 30 artículos duplicados, y los resultados de la búsqueda en *Google Scholar* presentaron un total de 21 documentos duplicados debido múltiples apariciones de los mismos artículos en diferentes idiomas y almacenados en diferentes fuentes bibliográficas o repositorios. De esta forma se obtuvo un total de 51 documentos descartados por duplicidad. Sobre el mismo proceso se rechazaron 291 documentos por aplicación directa de los criterios de exclusión e inclusión que, en adición a los 142 resultados descartados por no considerarse relevantes, sumaron un total de 434 elementos rechazados a partir de las búsquedas. Como resultado del proceso de filtrado sugerido se obtuvieron 13 documentos aceptados.

Como siguiente paso, se hizo un nuevo filtrado sobre la documentación aceptada definiendo una serie de criterios de calidad que se enumeran en la Tabla 2.8. Los criterios de calidad se definieron con base en las recomendaciones de [44] para manejar un enfoque desde la ingeniería y revisar los parámetros importantes de estudios que relacionan conceptos de *software* e información. Con esto se buscó evaluar la calidad de los documentos según la calificación de cada uno de los artículos filtrados anteriormente con base en cinco preguntas. Cada pregunta fue puntuada con los valores 1, 0.5 y 0, siendo equivalentes a las respuestas “Sí”, “Parcialmente” y “No” respectivamente. El valor máximo que un documento podía tener era de 5 y el umbral de selección considerado fue de 3. Los documentos por encima del umbral fueron seleccionados para esta investigación.

Pregunta	Puntaje
Q1. ¿Realiza una revisión comprensiva del estado actual del conocimiento o trabajos relacionados existentes?	Si / Parcialmente / No (1 / 0.5 / 0)
Q2. ¿Presenta un diagrama de los principales componentes arquitectónicos?	Si / Parcialmente / No (1 / 0.5 / 0)
Q3. ¿Muestra y describe los algoritmos o pseudocódigos utilizados o por lo menos indica claramente las técnicas utilizadas?	Si / Parcialmente / No (1 / 0.5 / 0)
Q4. ¿Analiza los principales hallazgos y resultados empíricos obtenidos?	Si / Parcialmente / No (1 / 0.5 / 0)
Q5. ¿Describe claramente la evaluación realizada de los resultados obtenidos?	Si / Parcialmente / No (1 / 0.5 / 0)

Tabla 2.8. Criterios de calidad. Fuente propia.

En la Tabla 2.9 se muestran los resultados obtenidos del proceso de evaluación de calidad de la documentación aceptada. Cómo se puede observar hubo un total de

4 artículos descartados al no alcanzar el margen de puntaje definido. Tras finalizar este proceso de filtrado se obtuvo un total de 9 artículos para ser usados en la presente investigación.

Título	Q1	Q2	Q3	Q4	Q5	Puntaje
Assisting researchers in bibliographic tasks: A new usable, real-time tool for analyzing bibliographies	1	0.5	0.5	1	1	4
Dialnet metrics as a bibliometric evaluation tool: Contributions to the analysis of the scientific activity in social sciences and humanities	1	0	0	0.5	0.5	2
An efficient author information retrieval tool for bibliographic record analysis	0.5	1	1	1	0.5	4
Software survey: ScientoPy, a scientometric tool for topics trend analysis in scientific publications	1	1	1	1	0.5	4.5
Bibliometrix: An R-tool for comprehensive science mapping analysis	1	1	1	1	0.5	4.5
SciMAT: A new science mapping analysis software tool	1	1	1	1	1	5
ScriptLattes: an open-source knowledge extraction system from the Lattes platform	1	0.5	1	1	0.5	4
LattesMiner: a multilingual DSL for information extraction from lattes platform	1	1	1	1	0	4
Um sistema para análise de redes de pesquisa baseado na Plataforma	1	1	0	0.5	0	2.5
Visual representation of bibliographic production data from Lattes Platform	1	0	1	1	1	4
Diseño de un sistema inteligente para estimación de categorización de grupos de investigación a partir de lineamientos definidos por COLCIENCIAS	1	1	1	1	1	5
Sistema de información web que procese y visualice los indicadores de investigación del Grupo Metis acorde a la normatividad de Colciencias	1	0	0.5	0.5	0.5	2.5
Desarrollo de una plataforma de migración de artículos publicados en GrupLAC a Scientific Journal Rankings (SJR)	1	0.5	0	1	0	2.5

Tabla 2.9*.* Resultados de la evaluación de calidad. Fuente propia.

2.2.5. Extracción de datos

Una vez seleccionados los documentos, se hizo una revisión más exhaustiva para captar los detalles relevantes de estos en relación con la temática planteada. Por

consiguiente, se realizó una lectura con mayor detalle para la extracción de los siguientes metadatos que complementaron a la posterior síntesis de la información y planteamiento de las brechas en el campo.

- **Re:** Referencia
- **AP:** Año de publicación
- **Te:** Temáticas
- **FB:** Fuentes de muestras bibliográficas
- **EA:** Elementos de análisis
- **TP:** Técnicas de procesamiento de datos
- **Vi:** Visualización

La Tabla 2.10 presenta de forma resumida y condensada la información principal de cada uno de los documentos seleccionados.

Re	AP	Te	FB	EA	TP	Vi
[45]	2021	-Bibliografía -Bibliometría	-Scopus -OpenCitations -CrossRef	-BTs o "*Bibliographic Tasks*" (BT1-BT13)	-Preprocesamiento -Procesos de *BIEs o "Bibliography Information Elements"*	-Línea de tiempo -Vistas narrativas -Barras e histogramas -Nube de palabras -Radar
[46]	2020	-Bibliografía -Bibliometría	-*DBLP o Digital Bibliography and Library Project*	-Desempeño	-Preprocesamiento -*DBLP partitioning*	-Grafos -Barras e histogramas
[47]	2019	-Cienciometría -Bibliometría	-Web of Science -Scopus	-Evolución -Desempeño -Tendencia	-Preprocesamiento -AVG o "*Average Growth Rate"* -ADY o "Average Documents per Year" -Etc.	-Línea de tiempo -Barras e histogramas -Gráfico de puntos -Nube de palabras
[48]	2017	-Bibliometría -Mapeo científico	-Web of Science -Scopus -Dimensions	-Redes -Evolución -Desempeño -*Burst detection* -Geoespacial	-Análisis descriptivo -Sondeo matricial -*Data reduction* -*Coupling* -*Co-occurrence* -Etc.	-Mapa factorial -Dendrograma -Mapa semántico -Red de colaboración -Historiografía -Etc.
[49]	2012	-Bibliometría	-Web of Science	-Redes -Evolución	-Preprocesamiento -*Network extraction*	-Diagrama estratégico

		-Mapeo científico	-Scopus	-Desempeño -*Burst detection* -Geoespacial	-*Clustering* -*Data reduction* -*Coupling* -Etc.	-*Cluster network* -Mapas superpuestos -Línea de tiempo -Etc.
[50]	2022	-Bibliometría -Mapeo científico	-Lattes	-Redes	-Preprocesamiento -Minería de texto -Proyección multidimensional	-Grafos -Barras e histogramas
[51]	2009	-Bibliometría -Minería de datos	-Lattes	-Redes -Geoespacial -Evolución	-Preprocesamiento -*Web scraping, HTML parser*	-Línea de tiempo -Red de colaboración -Radar -Gráfico de puntos -Mapa geográfico
[52]	2011	-Bibliometría -Minería de datos	-Lattes	-Redes -Geoespacial	-Estructuración de datos -*Web scraping, RegEx*	-Red de colaboración -Mapa geográfico
[53]	2019	-Inteligencia computacional	-GrupLAC	-Clasificación -Predicción -Recomendación	-Redes neuronales -*Web scraping, XPath*	-Línea de tiempo -Barras e histogramas

Tabla 2.10. Resumen de metadatos de los documentos relacionados. Fuente propia.

2.2.6. Síntesis

En esta sección se presentan los diversos trabajos relacionados con esta investigación, considerando las aproximaciones orientadas al flujo de datos bibliográficos de la actividad científica y sintetizando la información revisada. A través de este proceso, se conocieron las brechas existentes con respecto a la exploración y recuperación de datos, así como del análisis y visualización bibliométrica de los grupos de investigación en el Cauca. También se mencionan aquellas características que fueron tomadas como aportes para la presente investigación en el recorrido de cada trabajo revisado.

Existe gran variedad de investigaciones y herramientas propuestas para llevar a cabo las tareas de análisis y visualización bibliométrica en diferentes contextos. En primer lugar, se revisó el artículo [45] que presenta la herramienta "*VisualBib"*, una plataforma web creada con el propósito de proveer a los investigadores de una herramienta analítica y visual en tiempo real sobre la bibliografía científica de

fuentes globales. *VisualBib* permite explorar y analizar bibliografía de forma semántica y contextual usando metadatos y tipologías. La herramienta se enfoca en el acompañamiento y refinamiento de la creación de bibliografía para la generación de nuevos documentos. En consecuencia, no se concentra como tal en la revisión de métricas de impacto o desempeño. Sin embargo, *VisualBib* ofrece una cantidad enorme de funciones de análisis y visualización de datos bibliográficos de manera interactiva a través de una plataforma web, lo que se considera como una solución interesante en términos de arquitectura, diseño, narrativa, entre otras características.

Desde una perspectiva más general, en [46] se presenta una herramienta para la recuperación de información bibliográfica proveniente de *DBLP* (*Digital Bibliography and Library Project*). Se muestran algoritmos para la obtención y procesamiento de los datos con el fin de medir el desempeño de los autores y analizar sus perfiles. Para esto se hace uso de diversos indicadores bibliométricos como consistencia, factor de contribución, estabilidad, cooperación y solidez. A diferencia de las plataformas de *ScienTI*, los datos de *DBLP* pueden ser recuperados de forma estructurada en formatos como XML.

Por su parte, el artículo [47] realizado por investigadores de la Universidad del Cauca, presenta una herramienta cienciométrica, denominada "*ScientoPy",* para el análisis de temas que son tendencia en publicaciones científicas almacenadas en las bases de datos de *Scopus* y *Web of science*. La herramienta es capaz de desempeñar tareas cómo combinar datos de ambas fuentes, presentar indicadores*,* realizar análisis temporal y presentar opciones de visualización. *ScientoPy* posee un flujo de trabajo y algunos indicadores útiles cómo "*Average Growth Rate"* y "*Average Documents per Year"*. Sin embargo, la herramienta requiere de procesos manuales para la incorporación de los datos y se limita a dos fuentes globales de alto impacto, por lo que no resulta muy útil para el análisis del contexto regional. La presente investigación busca expandir este matiz de aplicación y automatizar procesos en el flujo de datos.

Entre las investigaciones más reconocidas de este tipo, se tiene el artículo [48] que presenta a "*Bibliometrix"*, una potente librería de código abierto desarrollada en *R* para análisis bibliométrico y mapeo científico. Consta de varias etapas como la recopilación, el análisis y la visualización de datos, terminando con la interpretación. Se enfoca en identificar una base de conocimiento y su estructura conceptual para

producir una red social de una comunidad científica para un tema en particular. Comparte semejanzas en el flujo de trabajo utilizado por [47], como la recopilación de datos cargados de forma manual, técnicas para el preprocesamiento de datos y otros procesos afines para su análisis. La herramienta aborda un vasto matiz de posibilidades de análisis y visualización, siendo una de las herramientas más completas para ejecutar análisis bibliométricos y presenta diversidad de opciones para su adaptación a otras soluciones.

Otra muy reconocida investigación es [49] que introduce a "*SciMAT"*, la cuál es una de las herramientas mejor diseñadas para el análisis de mapeo científico. Se caracteriza por ser un poderoso marco de referencia longitudinal que provee diferentes módulos para abordar etapas de recuperación de datos, preprocesado, extracción de red, normalización, mapeo, análisis, visualización e interpretación. La herramienta presenta opciones de visualización estadística y de redes basándose en una amplia gama de técnicas matemáticas y posee características similares a [47] en su etapa de entrada de datos. Los resultados de *SciMAT* se especializan en el mapeo científico y las redes de colaboración a grandes escalas para fuentes bibliográficas globales. Por su parte, este trabajo de grado busca enfocarse en la bibliometría fundamental sobre un territorio definido y sentar las bases de desarrollos de este tipo para su escalabilidad a redes entre diversos territorios. Por otra parte, existen características aprovechables en el flujo de trabajo de *SciMAT,* como también en su arquitectura, remarcando tanto diseños óptimos como técnicas y métricas avanzadas para un buen manejo de los datos.

En relación a las ideas anteriores, las características principales de algunas herramientas como [45] permiten incluir personal de poco o nulo conocimiento técnico y acceder de manera sencilla a la plataforma. Como valor agregado, se reconocen igualmente técnicas de preprocesamiento de datos útiles cómo la eliminación de duplicados, normalización de texto y etiquetado de categorías, también identificadas en [47], [48] y [49]. Algunos de estos estudios ejecutaron métodos de evaluación de sus herramientas a través de participantes y medidas comparativas teniendo en cuenta criterios de usabilidad, velocidad, facilidad y de más aspectos percibidos y valorados a través de experimentación. Esta metodología es reconocida como un modo apropiado de evaluación para herramientas de análisis bibliométrico.

Otras características destacadas se observaron en herramientas que incluyen operadores *API* entre sus interfaces para recuperar los datos en sus flujos de trabajo, particularmente [49] y [45]. Entre estas funciones se notó la opción de recuperar datos directamente de *Scopus* al ofrecer un medio de recolección de metadatos a través de algunas de sus *APIs* tales como "*Abstract Retrieval"* y "*Author Retrieval"*. Este es un método eficiente y confiable para complementar la información que está siendo analizada, dado el peso y prestigio de la mencionada fuente bibliográfica.

En adelante, se revisaron investigaciones de menor reconocimiento, aunque con un mayor grado de relación al alcance regional. En [50] se presenta un artículo acerca del procesamiento, análisis y visualización de datos de la plataforma *Lattes* de Brasil, la cual es perteneciente a la red *ScienTI* de donde provienen igualmente las plataformas *CVLAC* y *GrupLAC*. Como resultado se implementó una herramienta para explorar selectiva y visualmente datos de *Lattes* con respecto a grupos de investigación. Se destacaron las técnicas de minería de datos empleadas para obtener conocimiento a partir de simple texto plano y la exploración visual de datos multidimensionales. Se concluyó que es posible usar este enfoque para identificar cómo pueden estar relacionados los grupos de investigación unos con otros y evaluar sus datos bibliográficos. Sin embargo, algunos resultados fueron insatisfactorios debido a las limitaciones impuestas por el foco del estudio acotado a una baja cantidad de grupos pertenecientes a una misma institución. En el presente trabajo de investigación se busca acumular un mayor número de grupos de investigación pertenecientes a diversas instituciones.

Otros estudios centrados en la extracción de datos de *Lattes* pueden ser encontrados en [51] y [52], proponiendo herramientas de minería de datos desarrolladas por equipos que identificaron una problemática relacionada a la producción investigativa de un territorio y a la importancia de extraer conocimiento de estos datos. En [51] se reconoce la dificultad de extraer y resumir el conocimiento de utilidad proveniente de los datos de los grupos de investigación, puesto que no existía un medio eficiente para esta tarea más que la consulta individual mostrando dificultades similares a las del caso colombiano con la plataforma *GrupLAC*. Para esto se propone la solución "*ScriptLattes"* la cual consiste en un sistema de código abierto para la creación de reportes basados en hojas de vida de la base de datos de *Lattes.* La herramienta utiliza técnicas de minería de datos para extraer documentos *web* de los perfiles de grupos de investigación y tomar secciones

específicas para luego procesar y organizar su texto. También presenta reportes y gráficos usando los datos cuantitativos disponibles al final de su flujo de trabajo.
De forma muy similar a la propuesta anterior, la herramienta presentada en [52] de nombre "*LattesMiner"*, tiene la función de extraer de manera automática información a través de las hojas de vida de investigadores o perfiles de grupos de investigación presentes en las páginas *web* de *Lattes* para luego ser analizada y visualizada. *LattesMiner*, al igual que la herramienta *ScriptLattes* realiza un tratamiento de datos a partir de estructuras *HTML*. Estas herramientas fueron creadas hace años como una solución auxiliar necesaria. Sin embargo, exponen y abordan situaciones que se viven actualmente en el ámbito investigativo nacional con plataformas como *CVLAC* y *GrupLAC*. Sobre la misma retórica, se hace énfasis nuevamente sobre la necesidad y potencial de este tipo de soluciones

Finalmente, en [53] se presenta un trabajo de grado donde se implementa un sistema de recomendación de productos para los grupos de investigación inscritos en la plataforma *GrupLAC.* Se extrae la información de los productos realizando procesos de extracción de datos con una parte manual y otra de *web scrapping*. No obstante, los métodos usados para la extracción son poco convenientes debido a que operan sobre un conjunto exclusivo y predeterminado de grupos, es decir, no tienen en cuenta una gran población ni se implementa la automatización de estos procesos. El estudio realizado propone un sistema de recomendación utilizando redes neuronales y basándose en el modelo de medición de los grupos de investigación para sugerir actividades que conduzcan al escalamiento de categoría de dichos grupos. Para el desarrollo de su herramienta, se opta por utilizar el patrón arquitectónico modelo-vista-controlador, el cual es considerado para la implementación de algunos módulos de esta propuesta debido a que facilita el mantenimiento y la escalabilidad.

2.2.7. Brechas

Las secciones anteriores permitieron conocer investigaciones realizadas sobre el análisis de la actividad científica en diferentes contextos. Por un lado, algunas ofrecieron aportes a la presente investigación en materia de diseño, construcción y evaluación de soluciones, así como en el análisis del contexto manejado. Por otro lado, se identificaron algunas brechas existentes a abordar sobre la presente investigación. A continuación, se presentan las brechas encontradas con respecto sus principales elementos relacionados:

- **Etapa de extracción de datos:** Falta de un medio automatizado y eficiente para la obtención de datos relevantes, completos y actualizados sobre la actividad científica en el Cauca. No se encontró una herramienta efectiva para rastrear y extraer los datos bibliográficos generados por los actores del *SRCTI* del Cauca manejados en *CVLAC* y *GrupLAC*. La mayoría de las herramientas consideran solamente fuentes bibliográficas altamente reconocidas y divulgadas. Por otro lado, bases de datos globales, como *Scopus,* poseen herramientas de utilidad como sus *APIs* para extraer la información, aunque no directamente de los grupos de investigación. Sin embargo, esta información se articula parcialmente al contexto regional planteado, aunque impone sesgos ligados al contenido que es acotado para revistas de alto impacto.
- **Etapas de análisis y visualización bibliométrica:** Falta de una herramienta de análisis y visualización bibliométrica para el contexto de los grupos de investigación en el Cauca. No se encontró una herramienta a la vez adecuada y compatible con las fuentes de información y granularidad deseadas para la construcción de reportes de análisis y visualización bibliométrica de los grupos de investigación a nivel regional. Debe señalarse que es posible adaptar un flujo de trabajo a partir de la implementación de los principales elementos usados en este tipo de herramientas.
- **Algoritmos del flujo de datos:** Los algoritmos identificados para los diferentes procesos del flujo de datos en los sistemas revisados son demasiado puntales como para ser replicados sobre el enfoque de esta investigación. Aunque se trate del procesamiento de datos de la misma naturaleza, existe variación de fuentes, formatos, lenguajes de programación, sesgos y de más elementos enfocados a problemas específicos de los diferentes estudios. Algunos de los algoritmos incluso resultan no ser fácilmente accesibles o interpretables debido a su codificación final o a diversas restricciones. No obstante, fue posible considerar fragmentos en común para su parcial adaptación en tareas útiles de cálculo, limpieza, integración y manipulación sobre estructuras de datos.

2.3. MATERIALES Y MÉTODOS

En esta sección se presentan las herramientas y recursos en términos de *hardware* y *software* utilizados en este trabajo de investigación.

2.3.1. Hardware

Las herramientas *hardware* constan de dos computadores portátiles utilizados para llevar a cabo todas las etapas de tratamiento de datos, desde su extracción hasta su visualización. Los computadores portátiles son un *HP Envy 15* y un *HP 14 Notebook*. Se seleccionaron por sus características adecuadas para la codificación y ejecución de procesos del flujo de datos y desarrollo *web*:

- **"*HP ENVY 15 Notebook PC"*** con sistema operativo "*Windows 10 Home",* CPU "*Intel Core i7-4510U Quad-Core*" de 2 GHz, GPU "*NVIDIA GeForce 840M*" de 4 GB, Memoria RAM de 8 GB y disco duro de estado sólido con 512 GB de almacenamiento.
- **"*HP 14 Notebook PC*"** con sistema operativo "*Debian GNU/Linux 11 bullseye*", CPU "*Intel Core i3-3217u"* de 1.8Ghz, memoria RAM de 8GB y disco duro mecánico de 512 GB.

2.3.2. Software

El *software* seleccionado consta del lenguaje de programación necesario para desarrollar las diferentes etapas del sistema, los diferentes *frameworks* necesarios para el desarrollo de aplicaciones *web*, *DBMS* para la persistencia de datos, *ORM* de las bases de datos, herramienta de control de versiones, entornos de ejecución, entornos de despliegue y la variedad de librerías para la codificación. Se seleccionó el lenguaje de programación *Python* por su gran cantidad de librerías y *frameworks* disponibles para el tratamiento de datos, uso de *APIs* y el desarrollo de interfaces *web,* así como su adaptabilidad para la cohesión y acoplamiento con sistemas gestores de bases de datos y *ORM*, permitiendo desarrollar todo lo necesario para el sistema alrededor de un mismo lenguaje de programación de manera fluida. A continuación, se listan los elementos *software* esenciales utilizados en este trabajo de investigación:

Función	Software	Versión
Entornos de ejecución	Python	3.8.16
	Anaconda	2.3.2
	Jupyter Notebook	6.5.2
	Visual Studio Code	1.76.2
Control de versiones	Git	2.31.1
	Github	*Online*

Extracción de datos	Beautifulsoup4	4.11.2
	Pyquery	2.0.0
	Requests	2.25.1
	Selenium	4.8.2
	Regex	2022.10.31
	Lxml	4.9.1
	Affiliation Search API	*Online*
	Author Search API	*Online*
	Scopus Search API	*Online*
	Abstract Retrieval API	*Online*
	Author Retrieval API	*Online*
Procesamiento de datos	Numpy	1.24.2
	Pandas	1.5.3
	Openpyxl	3.0.10
	Unittest	3.8.16
Persistencia de datos	Pgadmin4	5.5
	PostgreSQL	13.4
	Psycopg2	2.9.3
	SQLAlchemy	2.9.3
Desarrollo web y Visualización de datos	Dash	2.8.1
	Flask	2.1.1
	Plotly	5.13.1
	Matplotlib	3.5.1
	NetworkX	2.7.1
	WTForms	3.0.1

Tabla 2.11. Software utilizado para el desarrollo del proyecto. Fuente propia.

Para los entornos de despliegue se consideraron los servicios de la Tabla 2.12 cuyo proceso de adaptación se explica en la sección 7.2 de la monografía.

Servicio	Características
Máquina virtual del departamento de telemática – Universidad del Cauca	8GB de RAM, procesador con 8 núcleos, 50GB de almacenamiento, y el sistema operativo Ubuntu Desktop 22.04.2 LTS.
Instancia t2.micro de Amazon Elastic Compute Cloud	1GB de RAM, 1GB de almacenamiento, 750 horas de uso con la capa gratuita.

Tabla 2.12. Servicios utilizados para el despliegue del proyecto. Fuente propia.

3. METODOLOGÍA

Para el desarrollo de este trabajo de grado, se empleó una metodología adaptada de *PMBOK* (*Project Management Body of Knowledge*) o el "Cuerpo de Conocimiento de Gestión de Proyectos", tomando la edición más actual de la guía *PMBOK* como elemento de referencia [54]. Esta guía profundiza en los principios de la dirección de proyectos y en los dominios de desempeño para la efectividad de las entregas. El proceso de adaptación realizado se encuentra descrito en el Anexo A. A continuación, se procede a condensar y definir los métodos y secuencias necesarios para la construcción de un sistema que permita extraer, persistir, analizar y visualizar datos bibliográficos en el ámbito de los grupos de investigación en el Cauca.

En primera instancia, se definió un **enfoque de desarrollo híbrido** para el proyecto con cualidades tanto adaptativas como predictivas debido a cierta incertidumbre inicial de cara a los requisitos, pues no se identificó la existencia de una herramienta que contemple las condiciones planteadas en este trabajo de investigación en sus diversas etapas. En segunda instancia, se le atribuyó **modularidad** al proyecto teniendo en cuenta los principios de bajo acoplamiento y alta cohesión para su diseño. Este fue dividido en 3 módulos:

- **Módulo Extractor *CvLAC-GrupLAC*:** Módulo encargado de la extracción y persistencia de datos provenientes de *CvLAC* y *GrupLAC*.
- **Módulo Extractor *Scopus*:** Módulo encargado de la extracción y persistencia de datos provenientes de *Scopus*.
- **Módulo *Dashboard*:** Módulo encargado del análisis y visualización bibliométrica de datos.

En tercera instancia, cada uno de los módulos propuestos se dotó de sus propias **fases de desarrollo** para su modelamiento, construcción y evaluación. Para esto, se definió un carácter **iterativo-incremental** dentro del enfoque escogido, de modo que cada módulo representa una iteración que se complementa con la anterior a manera de incrementos hasta obtener la funcionalidad completa. Las fases de desarrollo de los módulos se definieron por:

- **Fase de análisis:** Permite definir los requisitos del módulo y comprender los datos o la materia de análisis.
- **Fase de diseño:** Se encarga del modelado y representación del funcionamiento, basado en los requisitos.
- **Fase de codificación:** Consiste en la programación del diseño propuesto y su adaptación a los entornos considerados.
- **Fase de evaluación:** Consiste en la verificación y validación de los requisitos, o bien en la conducción de un juicio de valor.

En última instancia, se construyó el **ciclo de vida del proyecto**, resumido en la Figura 3.1. Con base en el ciclo de vida, se representó la **estructura de desglose del trabajo** mediante la definición y composición de los capítulos 4, 5 y 6 dentro de la estructura de la presente monografía. De este modo, cada uno de estos capítulos representa un paquete de trabajo diferente, conteniendo cada uno diversos subpaquetes equivalentes a las fases de desarrollo y a sus actividades desglosadas a manera de secciones y subsecciones.

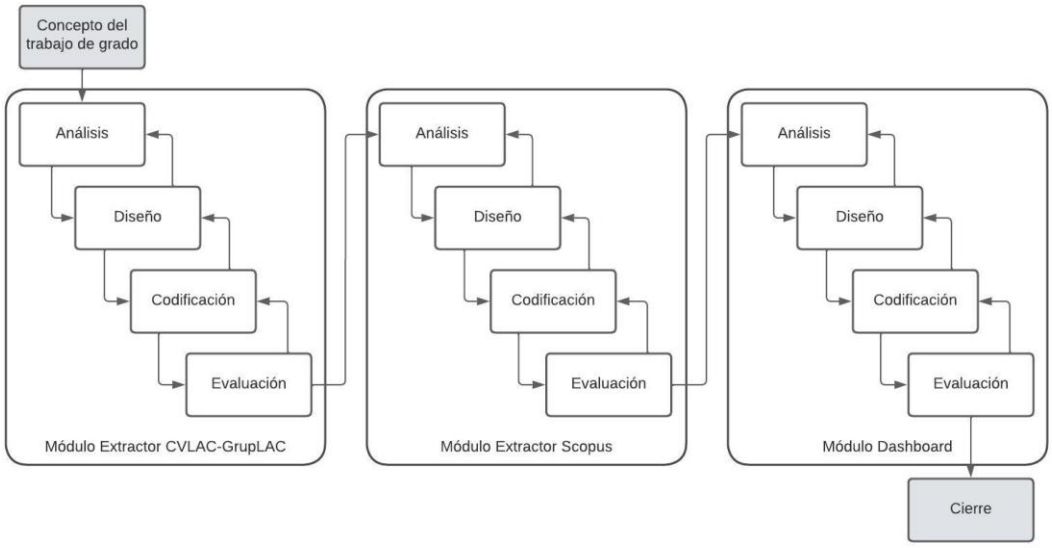


Figura 3.1. Ciclo de vida del proyecto. Adaptado de [54].

Para facilitar la lectura del trabajo realizado se estableció la documentación de la última fase del proyecto dentro del capítulo 7, esto debido a su importancia reflejada con respecto a la investigación en general. Finalmente, en los siguientes capítulos se presenta el desarrollo de la metodología planteada sobre el presente trabajo de investigación.

4. MÓDULO EXTRACTOR CVLAC - GRUPLAC

4.1. FASE DE ANÁLISIS

Para la identificación inicial de requisitos, se tuvo en cuenta a los actores interesados, el análisis de tareas, la lectura de documentos, la síntesis de la información, la estimación del alcance y las fuentes de datos bibliográficos estimadas. Cabe mencionar que la fase de análisis también se nutrió a partir del progreso, hallazgos y consideraciones particulares de las iteraciones incrementales de la metodología. El desarrollo de la fase comienza con la problemática y el escenario de motivación descritos sobre el contexto del *SRCTI* del Cauca y el papel de la bibliometría. Tras investigar, se identificó de qué manera se podía realizar un aporte aprovechando los datos disponibles y referentes a la actividad investigativa en el departamento. Se realizó una revisión sistemática de la literatura de la cual una parte fue dedicada a recolectar estudios realizados sobre fuentes bibliográficas latinoamericanas. Otra parte revisó los estudios realizados sobre la bibliometría y la investigación en general, acotando el contexto deseado y utilizando también bases de datos de revistas científicas internacionales y de alto impacto. Se encontraron las brechas de los trabajos relacionados y se plantearon los objetivos del presente trabajo de grado. Con esto se consolidó la base del conocimiento que permitió dar la orientación adecuada, de manera simultánea a la metodología aplicada. Finalmente, se tomaron decisiones para darle modularidad y flexibilidad al proyecto como respuesta a la complejidad e incertidumbre contingente. A continuación, se desarrolla a detalle la primera fase de análisis.

4.1.1. Análisis de fuentes de datos

Para este módulo, el tema definido se concentró sobre la extracción de datos de las fuentes disponibles en la red *ScienTI,* utilizadas de manera cotidiana por los investigadores y grupos de investigación a nivel nacional y regional: aplicativos *CVLAC* y *GrupLAC.*

Estos aplicativos se presentan a través de internet y rescatan la información de hojas de vida de investigadores y perfiles de los grupos de investigación por lo que ofrecen una gran variedad y cantidad de datos referentes a la investigación en el

territorio. Sin embargo, los perfiles sólo pueden ser observados individualmente y la visualización de la información en las dos plataformas se organiza a través de *DOMs* (*Document Object Model*), organizándola ambos aplicativos a través de tablas en una arquitectura *HTML* (*HyperText Markup Language*). Las tablas a su vez son apiladas una tras otra y están compuestas por filas y columnas. Esta información corresponde a los datos registrados por los usuarios y contemplados en los manuales de *CVLAC* y *GrupLAC* [14], [16] que derivan de alrededor de 80 tablas posibles por cada aplicativo.

Un ejemplo de un perfil de *GrupLAC* puede ser observado ingresando al enlace de [55] y uno de *CVLAC* en [56]. Cada tabla presenta información en particular, contenida en sus propias columnas (o características) y filas (o registros). Cabe mencionar que el orden de las tablas no se encuentra estructurado por completo en términos del código de la arquitectura de los aplicativos, por lo que muchos de los campos de información están concatenados en una misma etiqueta que debería corresponder a una única fila o columna. Un ejemplo de esto se puede observar en la Figura 4.1 con campos de información como país, fecha, título, páginas y más, mezclados en una misma cadena de texto. Esto demuestra la irregularidad con la que se presentan los datos y ciertas dificultades para su acceso.

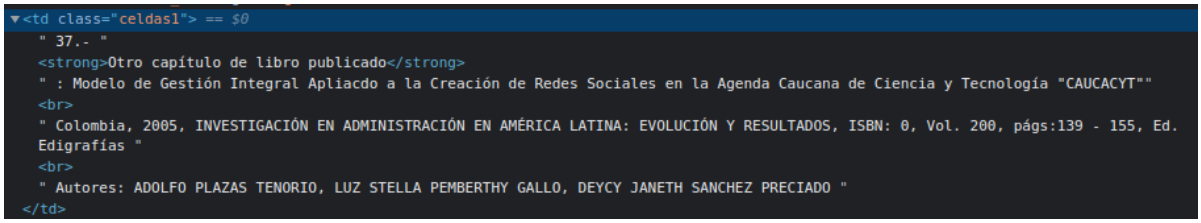


Figura 4.1. Fragmento de *DOM* de un perfil *GrupLAC***.** Tomado de [55]

Algunas de las tablas que corresponden a productos de investigación contienen marcas que simbolizan su validación en [21], el mensaje presentado por el aplicativo se observa en la Figura 4.2.

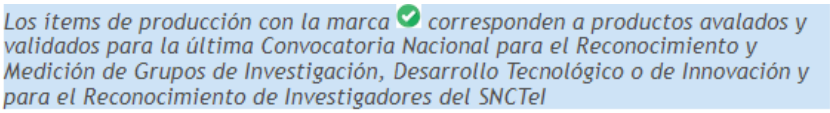


Figura 4.2. Mensaje de productos avalados por Minciencias**.** Tomado de [56]

El volumen de la información a nivel departamental se estima a partir de la totalidad de los grupos de investigación reconocidos por el Ministerio de Ciencia, Tecnología

e Innovación por una cifra de 118 grupos, incluyendo también a los investigadores miembros de estos grupos. Esta información se toma a partir de los servicios de búsqueda de Minciencias, que permiten observar los grupos de investigación por departamento, en este caso en el Cauca. [57]

4.1.2. Selección de los datos

La información disponible en las fuentes analizadas es abundante y excede el alcance limitado por el tiempo y la capacidad destinadas a la realización de este trabajo de grado. Por esta razón, se realizó una selección de tablas basándose, en un comienzo, en los "Pesos Globales de los Productos" que son resultado de los "procesos de investigación, desarrollo tecnológico e innovación" consignados en la última "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 - 2021" [21]. La intención fue priorizar las tablas que contienen información de mayor peso para el reconocimiento de un grupo de investigación teniendo en cuenta los tipos de producto principales. Los pesos designados permitieron asignar un criterio de selección para las tablas que son evidenciables en los perfiles de investigación. Las tablas de los perfiles pueden contener productos resultantes de varios subtipos, por lo que se incluyeron si al menos uno de ellos tenía un peso alto en comparación a los otros dentro de un mismo tipo.

Dado que se buscó un enfoque sobre los grupos de investigación, se prestó una atención selectiva a los elementos de los subtipos que resultan del grupo sobre los que resultan del individuo. Un ejemplo de esto es la dirección de tesis de doctorado que, aunque tiene un peso grande en su tipo, se prioriza como reconocimiento del individuo en su hoja de vida y normalmente no se registra en el perfil del grupo del sujeto si es que pertenece alguno. Por lo tanto, no resulta relevante para esta selección a pesar de su peso. También se da especial reconocimiento a productos que son usualmente utilizados en análisis bibliométrico según el estado actual del conocimiento como los artículos de investigación. Los pesos varían de 4 a 500 y algunos de los subtipos destacados inicialmente son:

- Artículos de investigación con Calidad A1 de peso 100
- Libros resultados de investigación con Calidad A1 de peso 300

- Productos tecnológicos con Patente de invención con Calidad A1 de peso 500
- Productos tecnológicos con Patente de modelo de utilidad con Calidad A1 de peso 500
- Empresas creativas y culturales con Calidad A de peso 100
- Innovaciones generadas en la gestión empresarial con Calidad A1 de peso 100
- Apoyo a creación de programas con Calidad A de peso 100

Posteriormente se tuvieron en cuenta intereses expresados por actores del *SRCTI* sugeridos por el asesor de este trabajo de grado y otros participantes del proyecto "*ECoS-CTeI"* en reuniones sostenidas. Igualmente, se tuvieron en cuenta tablas que presentan información básica, identificadores y datos generales sobre los investigadores o grupos de investigación. Estas últimas se seleccionaron a razón de su abundancia e importancia percibida en las hojas de vida y perfiles de los grupos, además de considerarse algunas de ellas como necesarios para el emparejamiento de los autores y los productos en etapas de integración y análisis de datos. Cabe mencionar que existen subtipos de productos de un peso grande con ausencia de producción en el Departamento del Cauca, por lo que terminaron siendo omitidos por escasez o nulidad de datos. La Tabla 4.1 y la Tabla 4.2 presentan las tablas de información de *CVLAC* y *GrupLAC* seleccionadas para la extracción de datos a partir de los criterios expresados anteriormente.

Nombre de tabla	Descripción	Selección
Información Básica	Presenta la información básica del investigador como nombre, nacionalidad, sexo, entre otros.	Datos necesarios
Identificadores de autor	Registro de identificadores de autor donde se encuentran registrados los usuarios (*WOS*, *Scopus*, *ORCID*, etc.).	Datos necesarios
Formación académica	Se refiere a etapas de educación para estudios generales o específicos de la formación profesional.	Interesados
Formación complementaria	Corresponde a los niveles de educación que complementan la formación académica del usuario.	Interesados
Áreas de actuación	Áreas de conocimiento en las cuales trabaja un investigador o presenta experiencias pasadas.	Interesados
Idiomas	Idiomas que maneja el usuario.	Interesados

Líneas de investigación	Identifica la información relaciona con las diferentes líneas de investigación en las cuales participa o ha participado el usuario.	Interesados
Reconocimientos	Reconocimientos o premios obtenidos por el usuario.	Interesados
Jurado en comités de evaluación	Actividades como jurado o de comisión evaluadora de trabajo de grado y tesis.	Interesados
Par evaluador	Participación como par evaluador.	Interesados
Artículos	Registra artículos de investigación publicados en revistas especializadas.	Peso e interesados
Libros	Registra algunos tipos de libros publicados principalmente libros resultado de investigación.	Peso e interesados
Capítulos de libro	Registra capítulos de libros resultado de investigación seleccionados por sus cualidades científicas como un aporte significativo a un área de conocimiento.	Interesados
Prototipos	Registra producciones técnicas y tecnológicas de prototipos desarrollados.	Peso e interesados
Redes sociales académicas	Contiene las redes sociales académicas donde se encuentran registrados los usuarios.	Datos necesarios
Empresas de base tecnológica	Registra producciones tecnológicas que son base de empresas o "*start ups".*	Peso e interesados
Softwares	Presenta producciones tecnológicas de programas de cómputo.	Peso e interesados
Innovación generada en la gestión empresarial	Registra producciones técnicas y tecnológicas de innovaciones generadas en la gestión empresarial.	Peso
Productos tecnológicos	Registra productos tecnológicos de varias categorías que pueden incluir patentes o secretos empresariales.	Peso
Estancias posdoctorales	Contiene las estancias posdoctorales en las que ha participado el usuario.	Interesados

Tabla 4.1. Tablas seleccionadas para la extracción de datos en *CVLAC.* Adaptado de [14], [21].

Nombre de tabla	Descripción	Selección
Datos básicos	Presenta la información básica del grupo de investigación como su clasificación, nombre, área de conocimiento, etc.	Datos necesarios
Instituciones	Contiene las entidades que ofrecen la infraestructura física y organizacional del grupo de investigación.	Datos necesarios

Líneas de investigación declaradas por el grupo	Contiene los enfoques interdisciplinarios que permiten englobar los procesos, prácticas y perspectivas de análisis. Sintetizan los estudios científicos y tecnológicos.	Interesados
Integrantes del grupo	Registra los miembros o investigadores del grupo de investigación y los enlaces a sus hojas de vida.	Datos necesarios
Programa académico de doctorado	Contiene los programas académicos de doctorado en los cuales ha contribuido el grupo para su creación.	Peso e interesados
Programa académico de maestría	Contiene los programas académicos de maestría en los cuales ha contribuido el grupo para su creación.	Peso e interesados
Otro programa académico	Contiene otros productos de formación y extensión en los cuales ha contribuido el grupo para su creación.	Interesados
Cursos de doctorado	Lista de cursos de doctorado diseñados para los programas nacionales y que contribuyen a las actividades del grupo.	Peso
Cursos de maestría	Lista de cursos de maestría diseñados para los programas nacionales y que contribuyen a las actividades del grupo.	Peso
Artículos publicados	Contiene artículos de investigación publicados en revistas especializadas.	Peso e interesados
Libros publicados	Contiene libros resultado de investigación publicados.	Peso e interesados
Capítulos de libro publicados	Registra capítulos de libros resultado de investigación seleccionados por sus cualidades científicas como un aporte significativo a un área de conocimiento.	Interesados
Otros artículos publicados	Contiene otros artículos de investigación publicados en revistas de menor reconocimiento. Regularmente son categorizados como notas científicas o *erratum*.	Datos necesarios
Otros libros publicados	Contiene libros publicados bajo categorías como libros de formación, libros de divulgación, entre otros.	Peso e interesados
Diseños industriales	Registra productos técnicos y tecnológicos de varias categorías que pueden incluir patentes o secretos empresariales.	Interesados
Innovaciones generadas en la Gestión Empresarial	Registra producciones técnicas y tecnológicas de innovaciones generadas en la gestión empresarial.	Peso
Plantas piloto	Contiene información acerca de un proceso físico o químico, mediante el diseño y construcción a escala reducida.	Peso e interesados
Otros productos tecnológicos	Registra productos tecnológicos de varias categorías que pueden incluir patentes o secretos empresariales.	Peso e interesados
Prototipos	Registra producciones técnicas y tecnológicas de prototipos desarrollados.	Peso e interesados

Softwares	Presenta producciones tecnológicas de programas de cómputo.	Peso e interesados
Empresas de base tecnológica	Registra producciones tecnológicas que son base de empresas o "*start ups".*	Peso e interesados

Tabla 4.2. Tablas seleccionadas para la extracción de datos en *GrupLAC.* Adaptado de [16], [21]

4.1.3. Recolección de los datos

La recolección de los datos se llevó a cabo a partir de la identificación del volumen de datos a nivel departamental mencionado en el análisis de las fuentes de datos. A partir de aquí se conoció la totalidad de los grupos de investigación que definieron el foco de la extracción. Recorriendo cada grupo de investigación en sus perfiles de *GrupLAC*, se pudieron encontrar los miembros pertenecientes a este y acceder a sus hojas de vida en *CVLAC*. Los datos se recolectaron a partir de la iteración de la lista de grupos y la lista de miembros de cada grupo, considerando las tablas seleccionadas y persistiendo los datos en bases de datos para su adecuada disposición. Para hacer esto posible fue necesario contar con funciones para extraer individualmente las hojas de vida y los perfiles de los grupos. La lista general de grupos en el Cauca fue consultada en la página *web* del buscador de grupos Minciencias por departamento [57], tomando, en cada solicitud, una versión actualizada de esta para la posterior iteración de sus elementos.

Uno de los puntos a considerar en la recolección de datos fue la posibilidad de actualizarlos y así fortalecer la fiabilidad de la herramienta y del proyecto. Por lo tanto, se abrió la posibilidad de hacer actualizaciones completas de los datos por parte de un usuario administrador o con suficiente conocimiento técnico para ejecutar la instrucción. Este proceso puede tomar mucho tiempo y limita el acceso a usuarios sin conocimiento técnico que podrían, al mismo tiempo, ser interesados o participantes del contexto investigativo. Por esta razón se abrió también la posibilidad de hacer pequeñas actualizaciones bajo demanda por parte de los usuarios sin conocimiento técnico, pudiendo extraer hojas de vida o perfiles de investigación de manera selectiva por un medio cómodo y fácil de usar.

Durante la recolección de datos, se aprovechó la identificación de códigos únicos para las hojas de vida y para los perfiles de los grupos que son tomados de los enlaces o *URL (Uniform Resource Locator)* de los aplicativos. Estos códigos fueron sumamente útiles para reconocer y ordenar los datos. A continuación, se observan

dos ejemplos de códigos *CVLAC* y *GrupLAC* respectivamente, definidos como "***idcvlac***" e "***idgruplac***":

*https://scienti.minciencias.gov.co/CvLAC/visualizador/generarCurriculoCv.do?cod_rh=**0001239368***

*https://scienti.minciencias.gov.co/gruplac/jsp/visualiza/visualizagr.jsp?nro=**00000000008160***

4.1.4. Definición de requisitos

Los requisitos sugeridos a partir de la fase de análisis para el Módulo Extractor *CVLAC-GrupLAC* fueron:

1. Extraer las tablas seleccionadas para todos los grupos de investigación del Departamento del Cauca y sus investigadores.
2. Extraer las tablas seleccionadas para un perfil *CVLAC.*
3. Extraer las tablas seleccionadas para un perfil *GrupLAC.*
4. Extraer las tablas seleccionadas para los perfiles *CVLAC* de los integrantes de un GrupLAC.
5. Persistir los datos extraídos en las bases de datos.
6. Mantener el nivel de granularidad y la integridad original de los datos desde sus fuentes.
7. Hacer un correcto manejo de excepciones y notificaciones durante la extracción.
8. Asistir a los usuarios con poco o nulo conocimiento técnico en tareas sencillas de extracción.

4.2. FASE DE DISEÑO

4.2.1. Definición de arquitectura

La arquitectura definida para el subsistema sigue el patrón arquitectónico *MVC* (Modelo Vista Controlador), el cual permite separar la lógica de funcionamiento de la interfaz del usuario y facilita la funcionalidad, mantenibilidad y escalabilidad del subsistema. En la Figura 4.3 se observa la arquitectura definida y las tecnologías utilizadas para su funcionamiento.

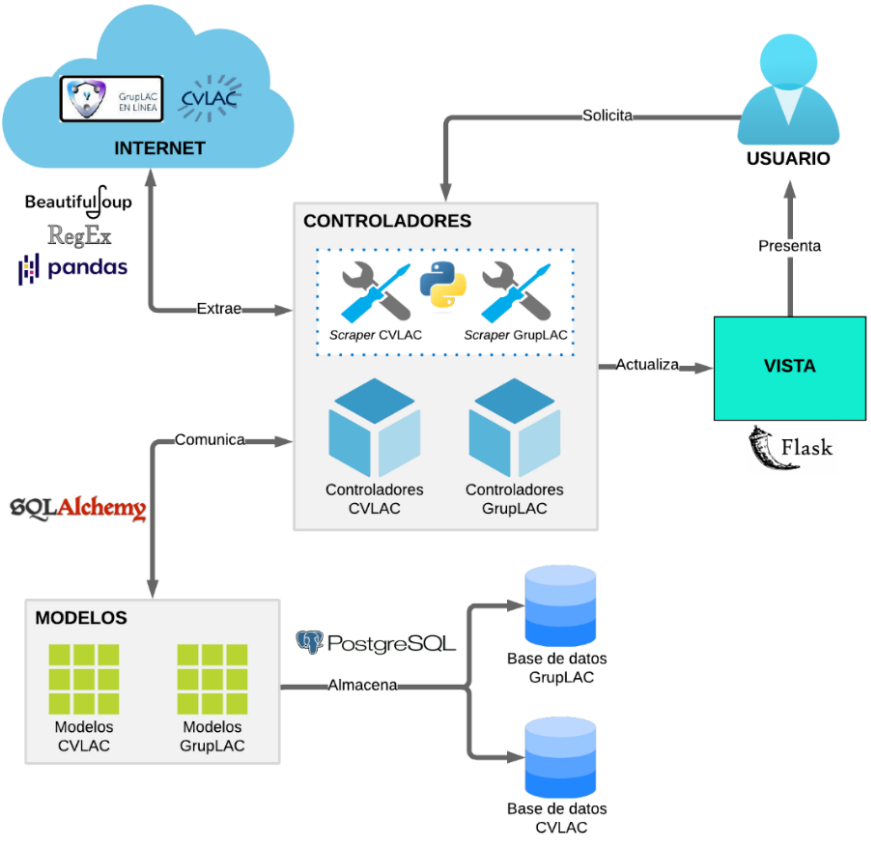


Figura 4.3. Arquitectura del Módulo Extractor *CVLAC-GrupLAC*. Fuente propia.

Los modelos son los encargados de acceder a las tablas de información y de representar los datos que serán almacenados o consultados. Estos actúan como intermediarias con las bases de datos y cada modelo representa una tabla perteneciente a una base de datos. Para este módulo se definieron dos bases de datos, una para *CVLAC* y otra para *GrupLAC*. Los controladores manejan las interacciones del usuario, solicitan datos a los "*scrapers"* o extractores, procesan estos datos y llevan a cabo las tareas de persistencia de datos con ayuda de los modelos. Al igual que los modelos, cada uno de los controladores corresponde a una de las tablas de las bases de datos e interactúan específicamente con el modelo encargado de esa tabla. La vista recibe reportes de los controladores tras cumplir sus tareas y los muestran al usuario de manera humanamente legible. La vista está diseñada para ser utilizada por un usuario sin conocimiento técnico, esto debido a que la función del usuario administrador se lleva a cabo mediante la ejecución de

scripts en el código fuente, dado que son tareas pesadas y tardías que comprometen la integridad del subsistema y requieren de cierta habilidad mínima.

Las tecnologías utilizadas se escogieron con base en su robustez, escalabilidad y soporte al ser cada una de ellas de tipo "*open source"* y contar con comunidades de usuarios que aportan documentación, mejoras, discusión, etc. Estas tecnologías proporcionaron herramientas de control en tareas de ingeniería de datos, procesamiento de datos y desarrollo *web*. Las tecnologías principales son brevemente descritas a continuación:

- ***Flask*:** Es un *microframework* web para *Python* útil para construir aplicaciones web pequeñas y medianas. Proporciona herramientas y bibliotecas para simplificar el proceso de desarrollo bajo una arquitectura modular y flexible. [58]
- ***PostgreSQL*:** Es un poderoso *DBMS* con más de 35 años de desarrollo activo con el cuál se ha ganado una fuerte reputación por su fiabilidad, robustez y desempeño. [59]
- ***SQLAlchemy ORM*:** Es el componente más famoso de la herramienta *SQLAlchemy* y permite mapear las clases de un programa en una base de datos de modo que el objeto de un modelo y el esquema de una base de datos se correspondan. [60]
- ***Beautiful Soup*:** Es una librería de *Python* para la extracción de datos de archivos *HTML* y *XML*. Provee de un gran número de elementos útiles para prácticas de *web scraping.* [61]
- ***Pandas*:** Es una librería de Python que provee estructuras de datos y herramientas para el análisis de datos de alto desempeño. [62]
- ***Regex - re*:** Regex (*Regular Expressions*) o expresiones regulares son patrones usados para el emparejamiento avanzado de combinaciones de caracteres en cadenas de texto. En *Python* se usa la librería "*re"* la cuál provee bibliotecas y operaciones aptas para su implementación. [63]

4.2.2. Diagramas de diseño

Con base en la arquitectura, se definieron los diagramas de diseño del módulo. Para la extracción de los datos se optó por seguir una programación orientada a objetos, las entidades y sus relaciones se observan en el diagrama de clases de la Figura

4.4, donde se presentan las clases, atributos y métodos resumidos, debido a la extensa cantidad de estos.

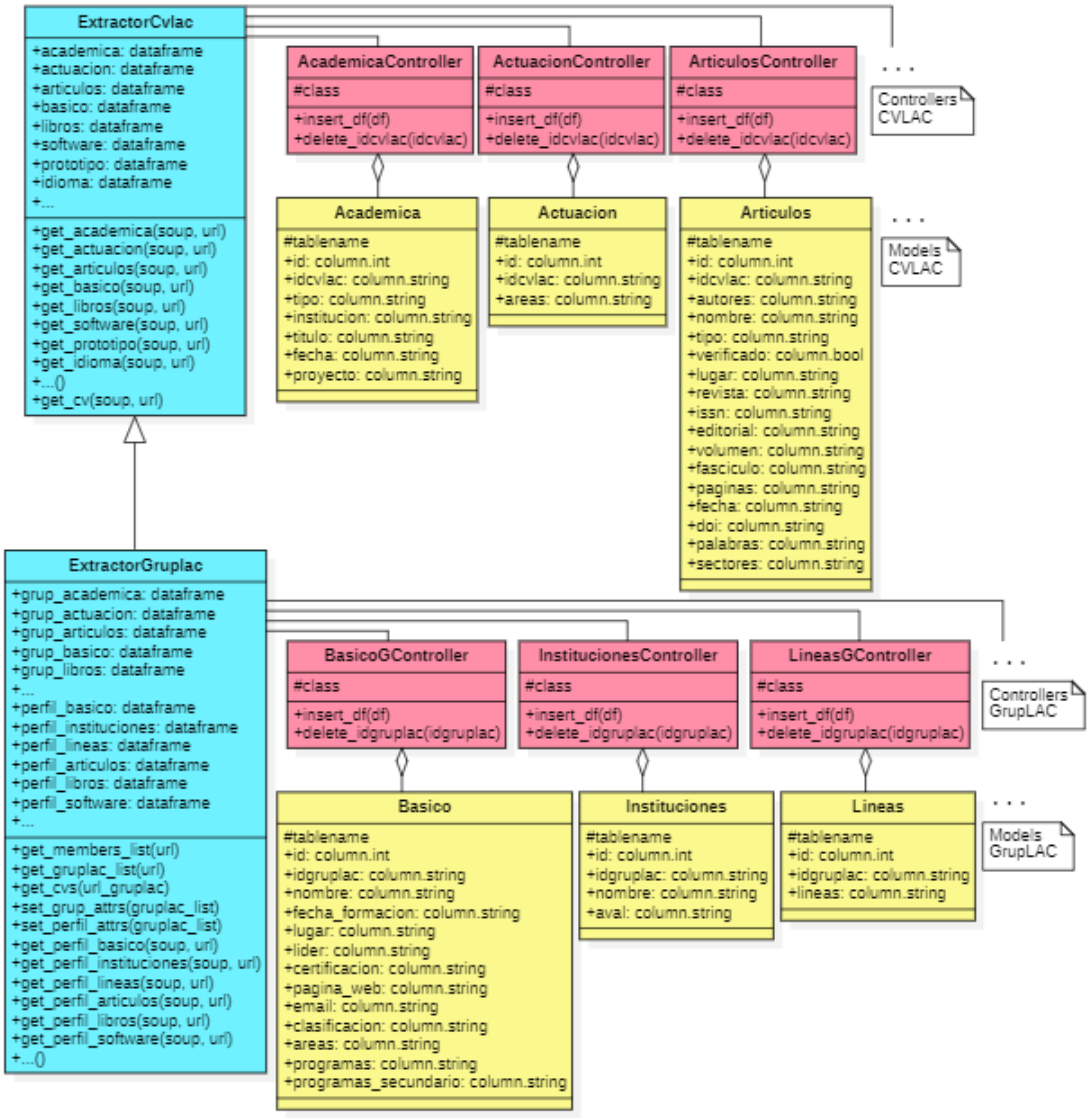


Figura 4.4. Diagrama de clases del Módulo Extractor *CVLAC-GrupLAC*. Fuente propia.

Los atributos de las clases extractoras corresponden a las tablas de información de las fuentes de datos, por lo que contienen las características únicas o columnas de estas, así como sus registros o filas. Sus métodos contienen a los algoritmos únicos a ejecutar sobre las tablas representadas por los atributos. Dado que es una cantidad extensa de tablas, varios de los atributos y métodos se omiten en el dibujo del diagrama para evitar saturarlo. Esto mismo se aplicó para la elaboración de los

controladores y modelos en el diagrama, entendiendo la correspondencia de estos a los elementos de la Tabla 4.1 y la Tabla 4.2, donde cada par controlador-modelo representa a uno de estos elementos para definir y operar las tablas de las bases de datos. En el diagrama de la Figura 4.4 se relacionan las siguientes entidades:

- **ExtractorCvlac:** Clase que contiene las variables y algoritmos correspondientes a las tablas de información de las hojas de vida de *CVLAC*. Se encarga de extraer y estructurar la información de una hoja de vida para hacerla manejable. Sus atributos corresponden a cada una de las tablas planteadas teniendo cada atributo la totalidad de columnas y registros de una tabla en una hoja de vida en particular. Por cada atributo existe un método único para el trabajo de extracción de su tabla en internet. Adicionalmente se tiene el método "*get_cv"* cuya función es usar a las otras funciones para extraer una hoja de vida completa.
- **ExtractorGruplac:** Hereda las características anteriores y contiene las variables y algoritmos correspondientes a las tablas de información de los perfiles de *GrupLAC*. Esta clase presenta el doble de atributos propios que su clase padre, diferenciados por sus prefijos "*grup"* y "*perfil"*. Cada prefijo tiene asignadas, una por una, todas las tablas de información seleccionadas para *GrupLAC*. El prefijo "*grup"* acompaña a las tablas correspondientes a las hojas de vida en *CVLAC* de los miembros de un mismo grupo de investigación, utilizando y extendiendo los atributos de la clase padre para acumular un conjunto de hojas de vida de un mismo origen en estos atributos propios de la clase hija. En otras palabras, los atributos de prefijo "*grup"* contienen un grupo de hojas de vida con una semántica en común. Ahora bien, los atributos de prefijo "*perfil"* emparejan las tablas de información propias del grupo de investigación que son observables en el perfil de *GrupLAC*. Para cada atributo de prefijo "*perfil"* existe un método único que extrae su tabla de internet. Para los atributos de prefijo "*grup",* se aprovechan los métodos heredados y se crean algunos métodos nuevos, propios de la clase hija, para gestionar las iteraciones de los individuos del grupo durante su extracción.
- ***Controllers*:** Son clases que se encargan de interpretar las instrucciones relacionadas con los datos en persistencia y los datos en extracción, cumpliendo tareas en función de estos según se solicite. Estas tareas corresponden a combinaciones de las operaciones básicas que se aplican en bases de datos relacionales, teniendo cada controlador métodos

adaptados para la tabla de su correspondencia. Además, están en constante comunicación con los modelos y el usuario.

- ***Models*:** Son clases especiales que representan y manipulan las tablas de información almacenadas en las bases de datos, respetando sus formatos, magnitudes y relaciones. Son usadas por los controladores para recuperar, modificar o agregar información según sea el caso. Junto con los controladores, funcionan como instrumentos de un *ORM* que asocia ambas clases para ejecutar operaciones sobre las bases de datos de manera programable y eficiente.

Para asistir a los usuarios que no tienen conocimiento técnico, se planteó desarrollar una pequeña aplicación *web* con el fin de proveer una interfaz gráfica de uso sencillo que permitiera cumplir con tareas básicas de extracción de datos en pocos pasos. El diseño se presenta a continuación mediante un diagrama de casos de uso en la Figura 4.5, también se observa el rol del usuario administrador, quien debe tener un conocimiento mínimo en ejecución de *scripts* de *Python* para construir el proyecto y realizar la extracción masiva de datos de forma relativamente sencilla y completamente automatizada.

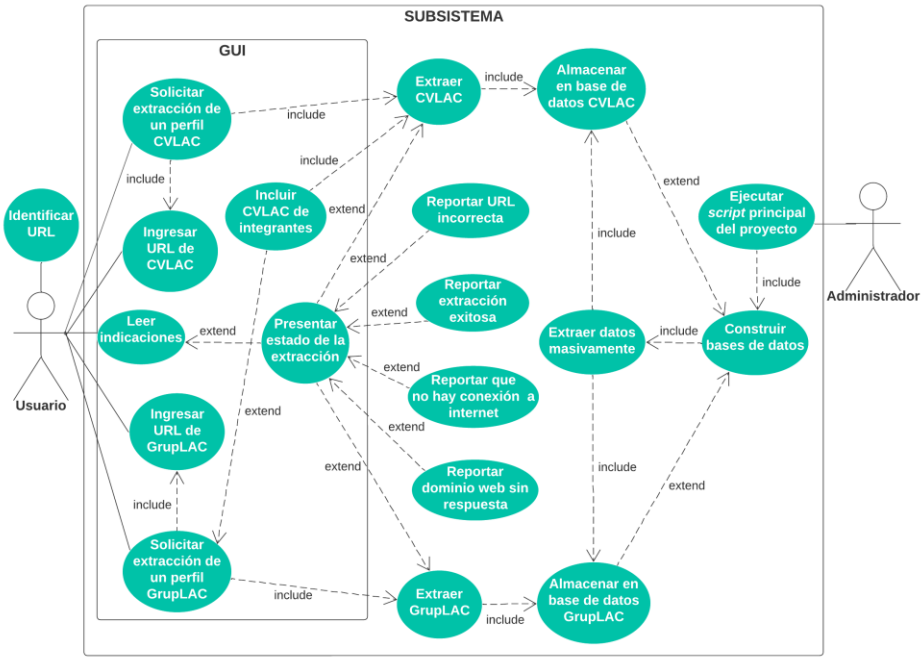


Figura 4.5. Diagrama de casos de uso del Extractor *CVLAC-GrupLAC*. Fuente propia

Las tablas almacenadas en las bases de datos se organizan de acuerdo con los diagramas entidad-relación presentados en la Figura 4.6, para *CVLAC*, y la Figura 4.7, para *GrupLAC*. En estos diagramas se pueden observar la totalidad de las tablas y sus columnas, las cuales son equivalentes a los valores internos de los atributos de las clases extractoras y las clases modelo del diseño presentado. La tabla que contiene la información básica de los investigadores, o del grupo de investigación, contiene valores únicos de sus códigos "*idclvac"* o "*idgruplac"* y se definen como llaves primarias. El resto de las tablas se relacionan por medio de llaves foráneas usando sus columnas de código de *CVLAC* o *GrupLAC* según sea el caso. Las tablas "*metadb"* son usadas para registrar la fecha en que se realiza la extracción masiva y la construcción de las bases de datos.

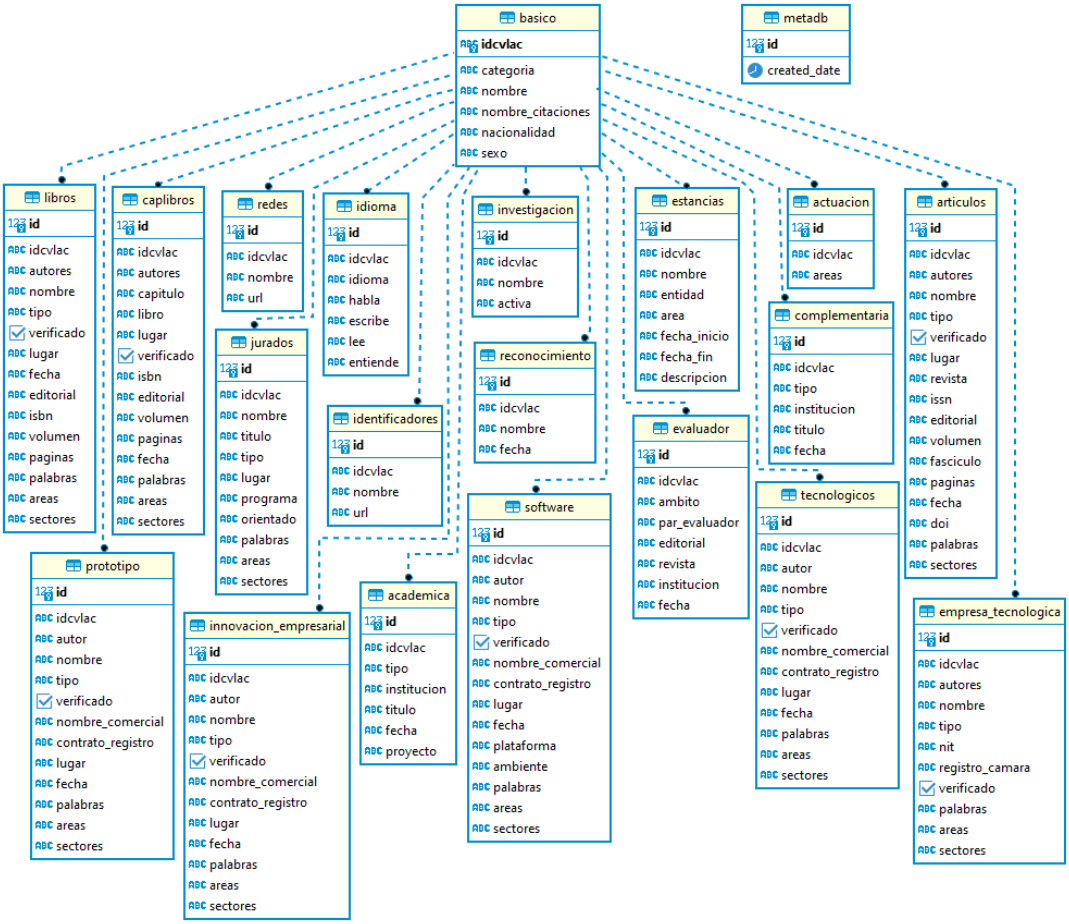


Figura 4.6. Diagrama entidad relación de la base de datos *CvLAC* del subsistema. Fuente propia.

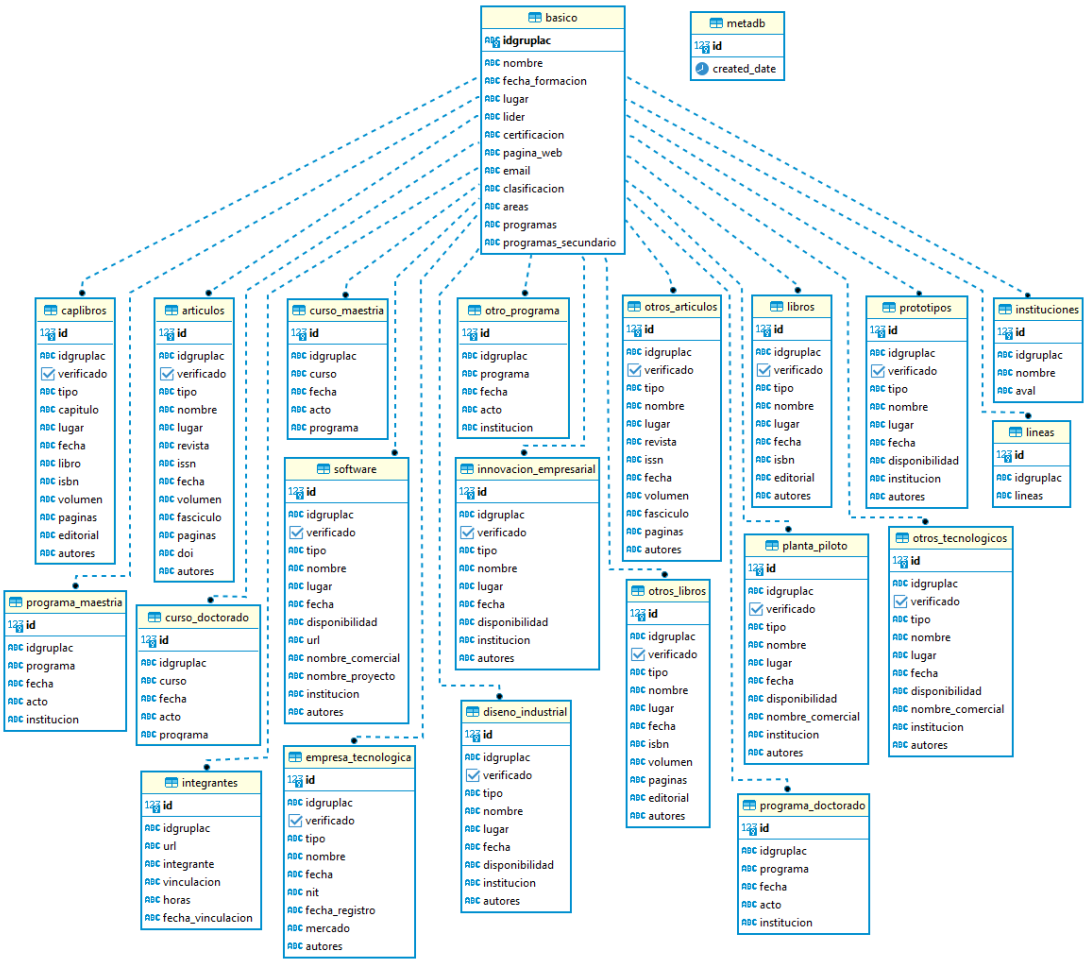


Figura 4.7. Diagrama entidad relación de la base de datos *GrupLAC* del subsistema. Fuente propia

4.3. FASE DE CODIFICACIÓN

El código desarrollado para este subsistema se puede encontrar en el repositorio compartido en el Anexo B de esta monografía, dentro del directorio "*cvlac*". Se debe leer el archivo *"README"* del repositorio para instalar el sistema del proyecto siguiendo los pasos indicados. A continuación, se describe el proceso de codificación realizado para este módulo.

4.3.1. Etapa de extracción de datos

Para la codificación de los algoritmos de extracción de datos de los aplicativos *CVLAC* y *GrupLAC* en internet, se utilizó una combinación única de técnicas de *web*

scraping, expresiones regulares y estructuras de datos de *Pandas* conocidas como "*dataframes*" en cada método. Siguiendo el diseño planteado, se comenzó por la programación de las clases extractoras. Los datos se toman de los documentos *DOM* de cada perfil de investigación a abordar. Los algoritmos o métodos implementados para las tablas son similares en sus entradas y salidas, aunque internamente podrían considerarse como cajas negras debido a su intrincada y particular programación. Estos algoritmos únicos abordan retos particulares de ingeniería de datos para identificar, extraer y organizar los datos de los aplicativos, así como el debido manejo de excepciones en cada caso.

Los procesos de extracción de hojas de vida y perfiles de grupos de investigación programados se encuentran especificados en los algoritmos 1, 2 y 3 descritos en el Anexo C de esta monografía. Hasta este punto, se construyeron las herramientas necesarias para realizar la extracción de muchos grupos de investigación y sus miembros. En la siguiente sección se explican los algoritmos de persistencia de los datos y su acoplamiento con los algoritmos de extracción para realizar extracciones masivas.

4.3.2. Etapa de persistencia de datos

La codificación necesaria para la persistencia de los datos consistió en programar los controladores y modelos mediante el uso de las herramientas del *ORM* y *DBMS* seleccionados. Lo primero en desarrollar fueron los modelos que representan las tablas de las dos bases de datos. Estos modelos se codificaron teniendo en cuenta las características de las tablas, los formatos de sus valores, las relaciones entre llaves, tamaño de los datos y propiedades de los campos de datos, entre otros protocolos exigidos por el *ORM*. De este modo, los modelos obtenidos pueden ser usados para construir las bases de datos rápidamente. Las bases de datos se definieron bajo los términos "*cvlacdb*" y "*gruplacdb*". Los modelos son complementados por los controladores que reciben estructuras de datos específicas desde la extracción, manejan sesiones y se soportan sobre funciones incorporadas del *ORM* para adaptar sus propios métodos.

Las funciones de persistencia personalizadas para este proyecto se construyeron, en su mayor parte, a partir de diversas funciones pre integradas de la caja de herramientas de *SQLAlchemy ORM*, las cuales asumen tareas complicadas con relación a las conexiones y compatibilidad con las bases de datos de *PostgreSQL*.

Los algoritmos de inserción, borrado, consulta y actualización de datos para las tablas de las bases de datos, así como el proceso para la extracción masiva a nivel del Cauca, se encuentran detallados en los algoritmos 4, 5 y 6 descritos en el Anexo C. Una vez programados los algoritmos, se realizaron varias extracciones de prueba a medida que se corrigieron errores, se manejaron excepciones y se implementaron numerosas técnicas para limpiar y mantener la integridad original de los datos en sus fuentes. A su vez, se practicó la persistencia y extracción masiva de datos para verificar su funcionamiento más adelante.

4.3.3. Interfaz Gráfica

La interfaz gráfica desarrollada ofrece las posibilidades de extraer una hoja de vida a partir de una *URL* de *CVLAC,* extraer un perfil de grupo de investigación a partir de una *URL* de *GrupLAC* y elegir la opción de extracción de las hojas de vida de los integrantes de un grupo de investigación. Las extracciones realizadas durante la sesión de un usuario actualizan las bases de datos que se construyen durante la extracción masiva de los datos, dado el caso en que los códigos identificados ya estén presentes, de otro modo se hace la inserción directa de los datos. Se trata de un tipo de actualización bajo demanda que se da por parte de los usuarios y consiste en el borrador de los datos relacionados a un código *CVLAC* o *GrupLAC* y su inmediata extracción y guardado.

La aplicación *web* notifica al usuario mediante mensajes sobre el estado de la operación que puede ser exitosa o presentar algún tipo de excepción relacionada con la conexión a internet, el estado de las plataformas de *ScienTI,* invalidez de la *URL*, entre otras. El usuario tiene la opción de desplegar un panel de instrucciones sobre el uso de la herramienta para guiarse en el proceso. A continuación, se presenta la Interfaz gráfica desarrollada en la Figura 4.8.

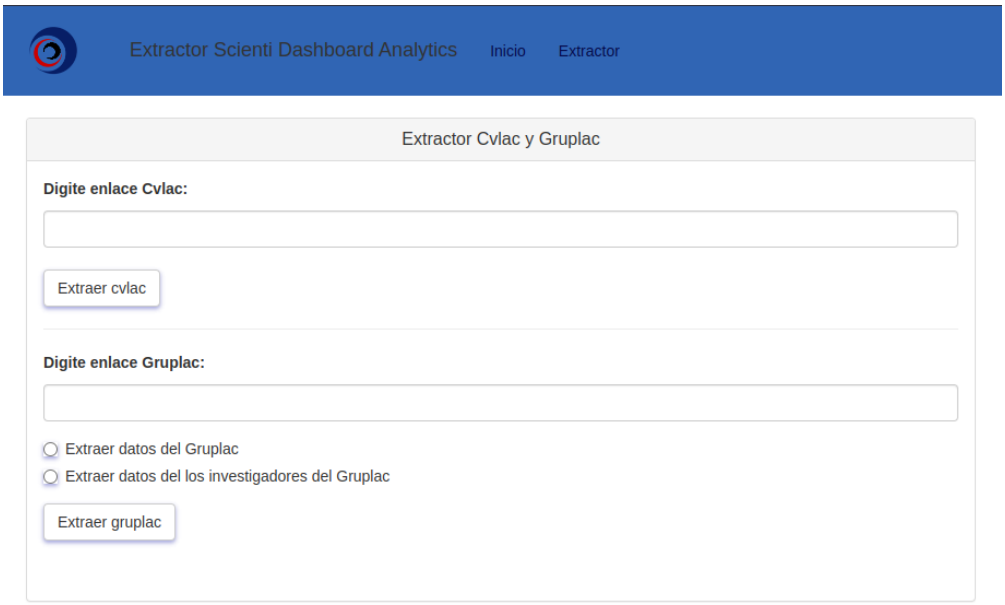


Figura 4.8. Interfaz gráfica de usuario del Módulo Extractor *CVLAC-GrupLAC*. Fuente propia.

Un administrador puede realizar la extracción masiva de los datos directamente desde el código del proyecto simplemente con ejecutar el *script* "***main.py***" presente en los recursos del repositorio compartido en el Anexo B para actualizar completamente las bases de datos. Sin embargo, este proceso puede ocupar varias horas según los recursos de cómputo y conexión a internet con los que se cuenten. Adicionalmente, la acción implica la limpieza total de los datos previos.

El subsistema permite la extensión del panorama propuesto dado que es posible extraer cualquier *CVLAC* y *GrupLAC* a nivel nacional y, del mismo modo, es posible asignar una *URL* que concentre los *GrupLAC* de otros departamentos a la extracción masiva, o bien de todo el país si se cuenta con la potencia computacional adecuada. No obstante, la presente investigación busca enfocarse sobre el Departamento del Cauca.

4.4. FASE DE EVALUACIÓN

4.4.1. Verificación

Para conocer si se construyó el subsistema correctamente, se comprobó el funcionamiento desde un punto de vista técnico. Esto se realizó mediante pruebas unitarias de "caja negra" para comparar los datos de entrada y los datos de salida

de todos los algoritmos de extracción de los perfiles de investigación. Se eligieron este tipo de pruebas debido a que estos algoritmos presentan una cantidad de detalles internos intrincada pero las características de sus entradas y salidas son bien conocidas y pueden ser evaluadas. Las pruebas unitarias se ejecutaron por separado para cada método del "*ExtractorCvlac"* y del "*ExtractorGruplac"* puesto que son la base del módulo extractor y son necesarios para su funcionamiento general. Las pruebas de verificación se programaron de modo que comparen estructuras de datos de tablas y arrojen el porcentaje de semejanza entre estas con base a la cantidad de cadenas de texto exactas entre cada celda, es decir, el conteo de campos de datos similares de las tablas entre pares de archivos.

Las tablas resultantes de los métodos únicos de los extractores, que se recuperan y procesan desde internet, se compararon con un conjunto de tablas extraídas manualmente desde perfiles de investigación tomados de manera aleatoria usando comandos *SQL* con la función "RANDOM()". Se le asignaron 3 *URLs,* al azar, a cada método de las tablas de *CVLAC* y *GrupLAC*. Por cada *URL,* se construyó manualmente la tabla especifica asignada para ese perfil en un archivo *CSV*, observando la página *web* y copiando digitalmente la información de cada columna y fila con el cursor. Este proceso se repitió a través de todos los métodos de las tablas seleccionadas, resultando en un total de 60 archivos *CSV* de extracción manual y 60 "*dataframes"* generados automáticamente solo para *CVLAC*. Para *GrupLAC* se tuvieron un total de 63 elementos multidimensionales manuales y 63 automáticos. Con todo listo, se organizaron los elementos tomados manualmente para ser comparados con los elementos que se generaron automáticamente desde los métodos.

La comparación se dio a través de las 123 pruebas unitarias programadas con la librería "*unittest*" de *Python*, 60 pruebas para *CVLAC* y 63 para *GrupLAC.* Cada tabla generada, automática o manual, contenía un número variable de hasta 18 columnas, dependiendo de sus características, y un número variable de filas, dependiendo de su cantidad de registros. Durante la ejecución de las pruebas de verificación, se reportaron a través de la consola del interprete los resultados de cada prueba, siendo este un porcentaje de diferencia o el término "*None"* para referirse a que no se encontró diferencia alguna. Cabe mencionar que esta etapa del trabajo de grado fue motivo de reiteraciones de las fases del módulo tras buscar los mejores resultados posibles con el tiempo disponible, considerando la enorme importancia que tiene la fiabilidad de los datos iniciales en proyectos de analítica y

bibliometría. En cada una de estas iteraciones se mejoró la limpieza de los datos y se variaron levemente los diseños y alcance del subsistema, hasta llegar a lo ya expuesto.

Los resultados de todas las pruebas de verificación para *CVLAC* y *GrupLAC*, ejecutadas por última vez a fecha de 7 de abril de 2023, fueron de "*None*", por lo que se asumió que el funcionamiento sobre las muestras tomadas fue correcto. Sin embargo, no se puede afirmar que el funcionamiento es perfecto pues el volumen total de datos abordados fue mucho mayor a las muestras. Además, la limpieza de datos implementada, aunque es robusta, no podría manejar la totalidad de errores posibles que pueden cometer los usuarios de las plataformas de *ScienTI* al digitar los datos de sus perfiles. No obstante, los resultados observados permitieron verificar que la información extraída es lo suficientemente confiable y fiel a la fuente incluso al alto nivel de detalle que se maneja, además de abordar correctamente los aspectos más críticos de la información, como el emparejamiento de las fuentes y la adecuada distribución de los datos granulados.

Se pueden ejecutar libremente las pruebas de verificación a través del proyecto del Anexo B usando el *script* ***"testing_main.py"*** como recurso de evaluación de la herramienta para posibles tareas de mantenimiento o monitoreo. Las observaciones hechas por el programa durante la verificación se presentan a través de la consola del interprete como se observa en la Figura 4.9, los resultados completos se encuentran en el repositorio del Anexo B siguiendo la ruta de directorios: "*cvlac*"→"*testing*"→**"*unit_test.txt*"**. Los archivos *CSV* extraídos manualmente, también se encuentran en el repositorio del Anexo B siguiendo la ruta de directorios: "*CvLAC*"→"*testing*"→**"*testing_cvlac*"** y **"*testing_gruplac*"**. Se debe tener en cuenta que estos archivos construidos manualmente contienen la información a la fecha mencionada y cualquier actualización futura en internet no se vería representada en estos.

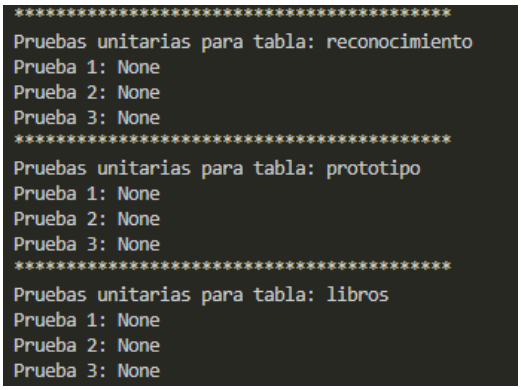


Figura 4.9. Fragmento del reporte de las pruebas de verificación para los *extractores CVLAC y GrupLAC*. Fuente propia.

4.4.2. Validación

Esta sección consistió en la validación de los requisitos planteados para este módulo con el fin de confirmar si se construyó el sistema adecuado. Por tanto, se buscó comprobar si los resultados cumplían con lo previsto en la fase de análisis. Para esto se pidió la asistencia de los investigadores Cristhian Nicolás Figueroa Martínez y Manuel Fernando Peláez Londoño, quienes realizaron la inspección de la herramienta para la primera y segunda validación (respectivamente) del cumplimiento de cada requisito. En la Tabla 4.9 se observa el resultado del proceso de validación. Las observaciones se tuvieron en cuenta para etapas subsecuentes del flujo de datos del sistema.

Requisito	Primera Validación	Segunda Validación	Observación
1	✓	✓	Evaluar la posibilidad de manejar excepciones de errores de instalación.
2	✓	✓	
3	✓	✓	
4	✓	✓	
5	✓	✓	
6	✓	✓	Evaluar la posibilidad de evitar duplicados.
7	✓	✓	
8	✓	✓	

Tabla 4.3. *Validación de requisitos del Módulo Extractor CVLAC-GrupLAC.* Fuente propia.

5. MÓDULO EXTRACTOR SCOPUS

5.1. FASE DE ANÁLISIS

Siguiendo la base de conocimiento, *Scopus* representa una exhaustiva base de datos bibliográfica examinada por expertos y enriquecida con la mejor literatura académica y científica alrededor del mundo. Como resultado, se ha posicionado como una de las fuentes de datos de investigación científica más importantes y utilizadas. El Módulo Extractor *Scopus* aprovecha esas facultades conectándose a *Scopus* a través de sus *APIs* para capturar la actividad científica registrada en el contexto deseado de manera similar al Módulo Extractor *CVLAC-GrupLAC*. Esto es posible mediante la suscripción de la Universidad del Cauca a esta base de datos, dado que otras fuentes consideradas, cómo *Web of Science,* imponen una mayor cantidad de limitantes y barreras de pago para el acceso eficiente a su información. Por este motivo y teniendo en cuenta el alcance, se optó por la alternativa más viable disponible con *Scopus*. Fueron necesarias varias solicitudes a la Universidad del Cauca y a la compañía *Elsevier* para obtener las credenciales y permisos necesarios para el desarrollo de este módulo y del proyecto en específico. A continuación, se desarrolla la Fase de Análisis del presente módulo.

5.1.1. Análisis de fuentes de datos

Las *APIs* contempladas en la Tabla 2.3 de esta monografía permiten explorar y extraer todos los datos bibliográficos de autores y productos registrados en la base de datos de *Scopus*. El total de herramientas disponibles para *Scopus* y otras plataformas de *Elsevier* se encuentran en [20], otras *API* permiten acceder a metadatos que no se consideran necesarios. El uso de estas herramientas está limitado a las cuotas y tiempos especificados. Como paso inicial, se identificó la necesidad de una "*API KEY"* para desbloquear las *APIs* necesarias*,* y de un "*Institutional Token*" para poder acceder a las interfaces fuera del instituto educativo. Estas dos credenciales fueron necesarias para el continuo acceso a los datos y, por lo tanto, para el correcto funcionamiento del módulo. Cabe mencionar que los miembros de las instituciones suscritas a *Scopus* pueden solicitar credenciales si el proyecto está justificado por las dos partes.

El proceso de solicitud para el acceso a los recursos mencionados se llevó a cabo mediante una serie de consultas a sectores administrativos de la Universidad del Cauca y solicitudes fundamentadas sobre los detalles de la presente investigación. Tras una serie de diálogos vía correo electrónico con la firma *Elsevier*, y con la asistencia del director y codirector, fueron aprobadas y compartidas las credenciales de acceso a las *APIs* de *Scopus*. Posteriormente, se observaron todos los datos disponibles para su selección y extracción a partir de las **vistas** de las *APIs* en: *Affiliation Search* [64], *Author Search* [65], *Scopus Search* [66], *Abstract Retrieval* [67] y *Author Retrieval* [68].

La fuente de datos permitió reconocer identificadores o códigos de autores y productos de investigación que también fueron registrados en algunos perfiles de investigación de las fuentes de *ScienTI*. Otros datos como *"DOI"* e "*ISBN*" facilitaron la identificación de artículos y libros, entre otros datos como su título. A través de las interfaces se pudo observar gran variedad de datos bibliográficos y estadísticos que resultaron útiles para el análisis, tales como la cantidad de citaciones de un producto o si este es de tipo "*open source"*. También datos relacionados con los autores y su actividad. La información se obtuvo mediante solicitudes construidas en *URLs* especificando parámetros y valores como se observa en el siguiente ejemplo:

*https://api.elsevier.com/content/author/**author_id**/57214098924?**view**=ENHANCED*

Los parámetros permiten seleccionar la información a partir de numerosas características descritas en [29]–[34], entre estas las afiliaciones, autores y productos. La respuesta retorna la información estructurada y organizada en formato *JSON* con el filtrado ordenado o bien un mensaje de error, según sea el caso. Un ejemplo se observa en la Figura 5.1. Todos los datos retornados provienen de la base de datos original de *Scopus* y se encuentran distribuidos en forma de llaves y valores.

Figura 5.1. Fragmento de respuesta para *Author Retrieval* en formato *JSON*. Fuente propia.

5.1.2. Selección de los datos

Las vistas disponibles de las *APIs* de *Scopus* contienen todos los datos posibles para la extracción y su descripción. Se buscó capturar la mayor cantidad de datos posibles con el fin de reconocer a profundidad a los autores y productos. Una mayor cantidad de características disponibles permitió un análisis más adecuado y abrió el abanico de posibilidades a futuro. Obtener cierta cantidad de dimensiones para los registros también permitió integrar estos datos con los que fue capaz de obtener el módulo del capítulo anterior, pues algunas variables pueden ser procesadas para realizar emparejamientos y abrir la posibilidad de distinguir los grupos de investigación en la base de datos que se genere. Las características seleccionadas se presentan a continuación, siendo omitidos campos restringidos, la mayoría de los enlaces web, variaciones textuales del mismo dato (se toma el dato original registrado) e historiales de afiliaciones de un elemento (se toman solo afiliaciones vigentes asociadas a ese elemento):

- ***Autores:*** Nombre preferido, nombre indizado, código de autor, código *eid*, código *orcid,* conteo de documentos publicados, fecha de creación de perfil, conteo de documentos citados, citaciones realizadas, *h-index,* conteo de coautores, estado del autor, áreas, rango de fechas de publicación, nombres de afiliaciones, códigos de afiliaciones, departamento o facultades de la afiliación. [68]

- ***Productos:*** Código *Scopus,* código *eid,* título, creador, nombre de la publicación, editorial, código *ISSN,* código *ISBN,* volumen, identificador *issue,* número de articulo, página de inicio, página de fin, conteo de páginas, fecha de publicación, idioma, código *DOI,* conteo de veces citado, enlace a vista previa, códigos de afiliaciones asociadas, resumen *abstract,* temas, tipo de fuente, tipo de documento, etapa de publicación, autores, códigos de autores, tipo de acceso, palabras clave de autor, palabras clave indizadas, nombres de afiliaciones, agencia fundadora, país. [67]

5.1.3. Recolección de los datos

Los datos se recolectaron construyendo solicitudes para las *APIs* de *Scopus* y personalizando los parámetros opcionales para segmentar la información que se retorna. Las *APIs Affiliation Search, Scopus Search* y *Author Search* son útiles para explorar la base de datos en busca de determinadas afiliaciones, productos o autores con base en sus códigos o identificadores, pero también por su territorio, campos de estudio, entre otros. Las *APIs Author Retrieval* y *Abstract Retrieval* retornan información con mayores niveles de detalle para complementar la información que proveen las anteriores. Dentro de este marco, los datos de la actividad científica a nivel del Cauca en *Scopus* presentan varias dificultades para su identificación puesto que las herramientas permiten buscar los elementos de un territorio únicamente por un nombre de ciudad o país, pero no por la extensión del territorio denominado como departamento en Colombia. Existen autores y productos en el Departamento del Cauca afiliados a instituciones que asocian su territorio de manera variada, por ejemplo, usando la cadena de texto “Cauca” en el campo del país, o cadenas de texto de los diferentes municipios del departamento en el campo de ciudad. Otras se asocian directamente a la ciudad de “Popayán” usando esta y otras variaciones resultantes de errores de registro al digitar la cadena, como por ejemplo “*popayán*” o ”*Popayan*”, entre muchas otras.

Por estas circunstancias, se recolectaron todas las afiliaciones a nivel del Departamento del Cauca realizando solicitudes a la base de datos con cadenas de texto normalizadas, sin tener en cuenta acentos ni mayúsculas, y usando parámetros personalizados para cada municipio del departamento y para la ciudad capital, entre otras palabras clave. Las solicitudes usadas para este proceso se encuentran en el Anexo D de esta monografía, junto con la cantidad de afiliaciones retornadas y descartadas en cada caso.

La selección de las afiliaciones se dio de manera minuciosa e individual de modo que, para cada afiliación retornada en cada grupo de afiliaciones observables de las respuestas, se realizó una búsqueda de su sitio *web,* o bien, de sus perfiles oficiales en internet a modo de verificación de su ubicación oficial. Cada afiliación perteneciente al Departamento del Cauca fue incluida a una lista de afiliaciones que fue iterada para encontrar a sus autores y productos derivados. Por último, se obtuvo una lista de afiliaciones depurada que se acomodó al contexto del presente trabajo de investigación y que fue usada como medio de recolección de los datos para el Módulo Extractor *Scopus*. No obstante, la extracción de autores o productos es posible si se cuenta con sus códigos, independientemente de si pertenecen o no al Cauca y según estime un usuario.

La extracción masiva de datos del Departamento del Cauca se da por acción de un usuario con conocimiento técnico, siendo el caso similar al módulo del capítulo 4, en donde se ofrecen herramientas para la actualización completa o parcial de los datos persistidos y la participación de distintos usuarios.

5.1.4. Definición de requisitos

Los requisitos definidos para el subsistema son:

1. Extraer las características seleccionadas para todos los investigadores y productos científicos a nivel del Departamento del Cauca en *Scopus*.
2. Extraer las características seleccionadas para cualquier autor en *Scopus.*
3. Extraer las características seleccionadas para cualquier producto en *Scopus.*
4. Persistir los datos extraídos en una base de datos.
5. Mantener el nivel de granularidad y la integridad original de los datos desde sus fuentes.
6. Hacer un correcto manejo de excepciones y notificaciones durante la extracción.
7. Asistir a los usuarios con poco o nulo conocimiento técnico en tareas sencillas de extracción.
8. Integrar los datos generados a partir de *Scopus* con los datos generados a partir del aplicativo *GrupLAC*.

5.2. FASE DE DISEÑO

5.2.1. Definición de arquitectura

El patrón arquitectónico *MVC* permitió definir la arquitectura del presente módulo de manera similar al módulo del anterior capitulo. En la Figura 5.2 se presenta la arquitectura definida. La extracción de los datos se realiza manejando las solicitudes y respuestas de las *APIs* a través de un extractor que contiene los métodos necesarios para comunicarse con estas y reestructurar la información que retornan. El módulo cuenta solamente con dos controladores y dos modelos para el manejo de datos de autores y productos, se sigue el mismo funcionamiento definido en el capítulo anterior en cuanto a las operaciones de persistencia sobre la base de datos y la comunicación simultanea con el extractor y el usuario. El módulo cuenta con una única base de datos que contiene un total de dos tablas representadas por los dos modelos, por lo tanto, una de las tablas contiene todas las características seleccionadas para los autores y la otra para los productos.

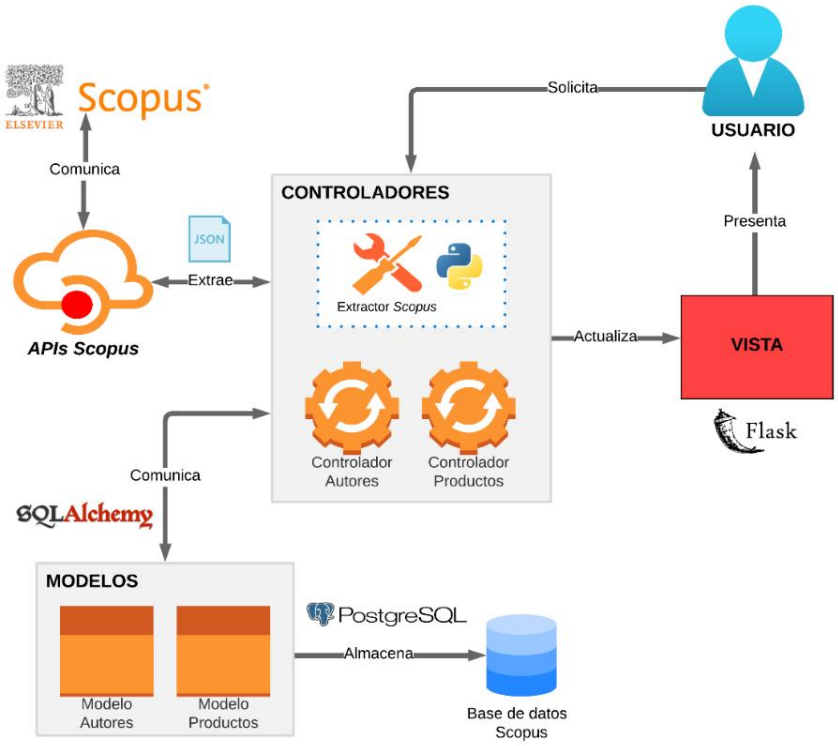


Figura 5.2. Arquitectura del Módulo Extractor *Scopus*. Fuente propia.

Los controladores reportan los estados de las tareas solicitadas por el usuario a través de la vista que se destina para usuarios sin conocimiento técnico. Un usuario con conocimiento técnico es capaz de ejecutar tareas más complejas con la simple ejecución de "*scripts"* y observar igualmente los reportes a través de una consola ordinaria*.* Se escogieron nuevamente las tecnologías utilizadas en el módulo anterior con excepción de las usadas para la extracción, puesto que la complejidad de estas tareas se ve considerablemente reducida en este módulo.

5.2.2. Diagramas de diseño

En la Figura 5.3 se presenta el diagrama de clases del módulo. La clase "ExtractorScopus" posee como atributos las tablas para autores y productos, además de dos variables que tienen asignadas las credenciales para el uso de las *APIs*. Por otro lado, se observan las clases asociadas de los controladores correspondientes a manejar las operaciones de cada tabla a través de sus modelos, por esta razón los atributos de las clases modelo corresponden a las características de las tablas. Los métodos de los controladores permiten insertar estructuras de datos o borrar todos los datos que coincidan con un código de afiliación o código de autor. El controlador "MetaDBScoController" y su modelo, sirven únicamente para llevar un registro en persistencia de la fecha en la que se hizo la última extracción masiva de datos en el módulo, por lo que también representa una tabla extra en la base de datos que se adiciona a las tablas originales de la información bibliográfica. El Módulo Extractor *CVLAC-GrupLAC* también presenta esta característica, no obstante, ninguna de las tablas, controladores y modelos de prefijo "*Meta"* en los dos módulos presentan funciones adicionales a la de registrar la fecha de ejecución de las extracciones. En cuanto a los métodos de la clase extractora, estos tienen el propósito de explorar, solicitar, organizar y disponer los datos que provienen de *Scopus*. Las funciones permiten desde la extracción de un autor o un producto, hasta la iteración de códigos de afiliación para obtener su actividad científica.

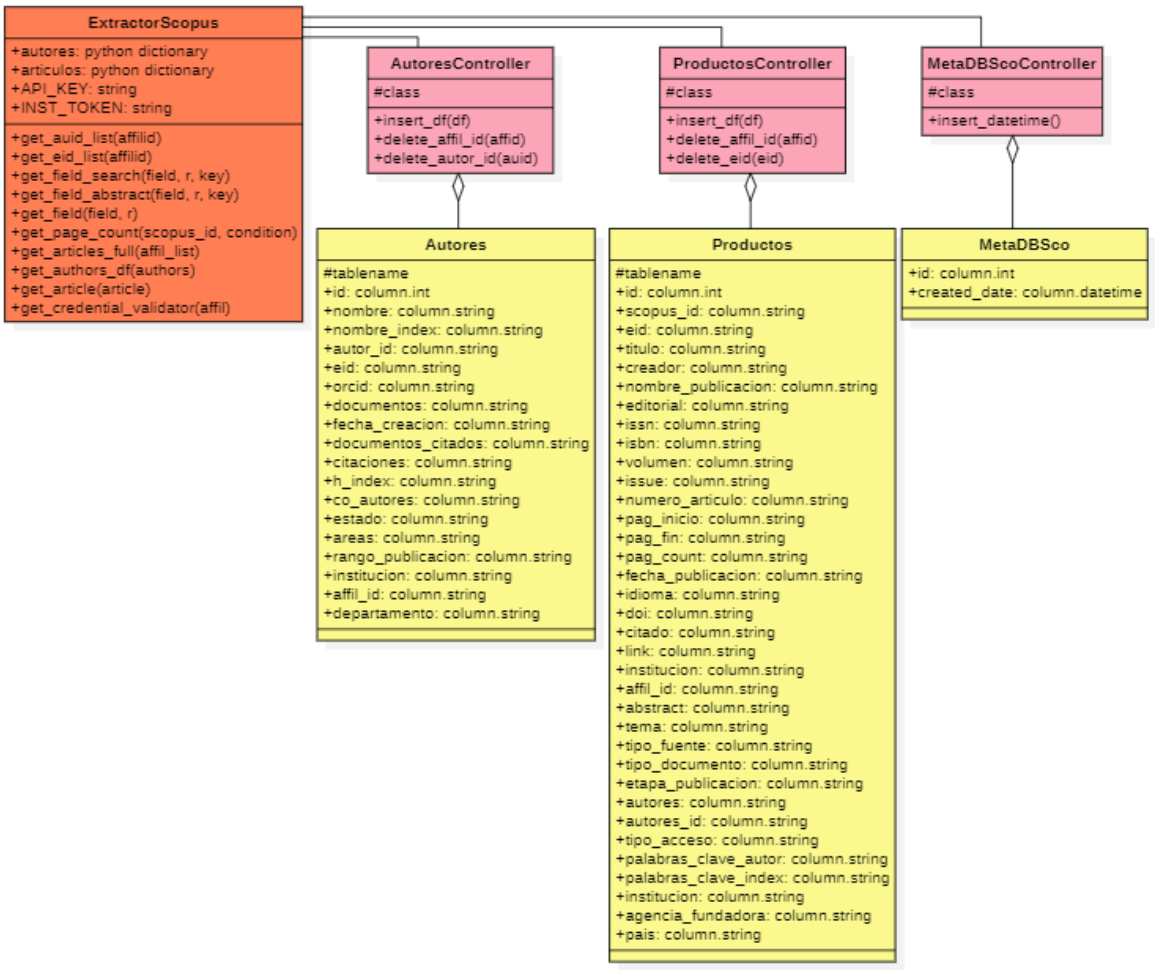


Figura 5.3. Diagrama de clases del Módulo Extractor *Scopus*. Fuente propia.

Para la ejecución de tareas sencillas de extracción por parte de usuarios sin conocimiento técnico se destinó una aplicación *web* simple para proveer una interfaz gráfica de usuario. En el diagrama de la Figura 5.4 se pueden observar los casos de uso del usuario y el administrador en su interacción con el subsistema.

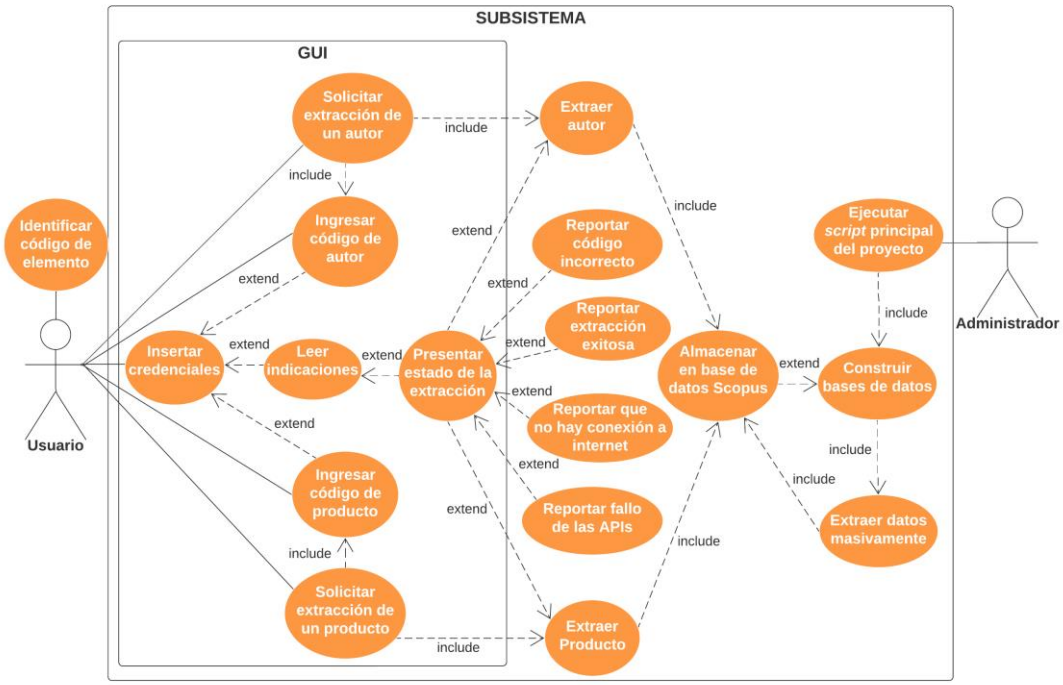


Figura 5.4. Diagrama de casos de uso del Extractor *Scopus*. Fuente propia.

Este módulo cuenta con una única base de datos de un total de 3 tablas como se observa en el diagrama de la Figura 5.5. Las tablas no tienen llaves foráneas dado que los códigos de identificación de afiliaciones o autores, que podrían usarse para este fin, no presentan valores únicos en sus registros al retornarse varios códigos concatenados para un mismo registro en diversos casos. Las condiciones en que las *APIs* de *Scopus* retornan la información no siempre se acomodan a los requisitos de estas relaciones, siendo que los segmentos de información extraídos pertenecen a entramados de conjuntos de datos inmensos de *Scopus* que extralimitan la información del contexto local. Por ejemplo, un registro individual de un artículo con multitud de identificadores de autor o de afiliación concatenados a modo de texto y pertenecientes a diferentes territorios internacionales, en una misma celda de información.

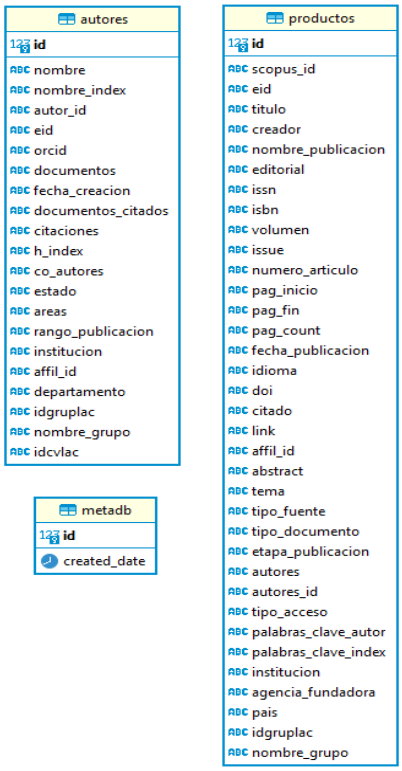


Figura 5.5. Diagrama entidad relación de la base de datos *Scopus* del subsistema. Fuente propia.

Los metadatos que configuran las relaciones en la base de datos original de *Scopus* exceden las capacidades de las *APIs* sobre las que está soportado este módulo, por lo tanto, son inaccesibles desde los parámetros de sus vistas que únicamente ofrecen datos bibliográficos e identificadores abiertos como se ha mencionado anteriormente. La información persistida en las tablas del presente módulo a su vez depende de información que ha sido recuperada de una base de datos mucho más abundante y compleja, que conecta información a nivel mundial. Sin embargo, la permanencia de códigos propios de *Scopus,* permitidos e identificables entre los productos y autores, fueron suficiente para realizar las tareas de emparejamiento internamente por medio de operaciones del *DBMS* y el *ORM.*

5.3. FASE DE CODIFICACIÓN

El código desarrollado para este subsistema se puede encontrar en el repositorio compartido del Anexo B de esta monografía, dentro del directorio “*scopus*”.

5.3.1. Etapa de extracción de datos

La clase “ExtractorScopus” contiene los métodos de extracción de características o campos de información de las tablas de autores y productos, estos son: “*get_field*”, ”*get_field_search*”, “*get_field_abstract*” y “*get_page_count*”. Dichos métodos exploran y capturan datos específicos contenidos en el archivo *JSON* retornado por las *APIs*. Otros métodos se encargan de formar listas de códigos a partir de afiliaciones, como “*get_auid_list*” (obtiene una lista de códigos de autor) y “*get_eid_list*” (obtiene una lista de códigos de productos). Los demás métodos presentan una mayor complejidad y utilizan a los métodos anteriores para construir grandes estructuras de datos que se almacenan en los atributos de la clase. La programación para la extracción de un autor y un producto, o de todos los autores y productos de una afiliación, se describe mediante los pseudocódigos de los Algoritmos 6 y 7 indicados en el Anexo C.

5.3.2. Etapa de persistencia de datos

La persistencia de los datos extraídos se da mediante la asistencia del *ORM* seleccionado cómo se apuntó en el capítulo anterior. Los procesos codificados para persistir los datos y para ejecutar la extracción masiva de datos a nivel del Cauca en *Scopus* se explican en el Algoritmo 9 indicado en el Anexo C. Los programas codificados se agregaron al “*script”* principal del proyecto “**main.py**” junto al proceso del código del módulo anterior para realizar la extracción a nivel del Cauca tanto en los aplicativos de *CVLAC* y *GrupLAC*, como de la base de datos de *Scopus.* La ejecución del proceso implica la limpieza y reconstrucción de la base de datos del presente módulo y es recomendado que sea ordenado por un usuario con mínimo conocimiento técnico.

5.3.3. Etapa de integración de datos

Los conjuntos de datos parcialmente generados por los módulos extractores, permitieron conocer las características en común entre los distintos elementos y, a través de estos, hacer una integración de los datos con el fin de reconocer a los grupos de investigación presentes en la base de datos generada a partir de *Scopus*. En este orden de ideas, los autores recolectados desde *Scopus* se emparejaron con los investigadores de *CVLAC* por medio del identificador de autor *Scopus* que un

investigador puede registrar en su hoja de vida. Posteriormente este investigador puede ser reconocido como miembro de un grupo de investigación en *GrupLAC* e identificar a qué grupos de investigación pertenecen los autores captados desde *Scopus*. De forma similar, los productos de los grupos de investigación, repartidos en numerosas tablas, contienen datos que se pueden identificar en las tablas del presente módulo. El *DOI* de los documentos, el *ISBN* de los libros y el texto normalizado de los títulos sirven como identificadores para la integración de los datos de ambos módulos.

El proceso realizado es presentado con mayor detalle en el Algoritmo 10 indicado en el Anexo C. Se debe tener en cuenta que los registros originales de los aplicativos *CVLAC* y *GrupLAC* no pasan por un proceso tan riguroso como sí lo hacen los registros de *Scopus,* resultando en dificultades para comparar los datos de texto por errores que resultan de digitarlos descuidadamente o anexar enlaces de internet incorrectos. Adicionalmente, existen registros de productos que no cuentan con los identificadores *DOI* o *ISBN,* que son el mejor medio para reconocer los productos entre los conjuntos de datos. Esta situación condiciona los datos y los expone a ser descartados en el proceso de integración si estos no se ven emparejados o reconocidos. El sesgo proviene mayormente de la calidad de los datos registrados en los aplicativos, es decir, por qué tan comprometidos y precisos son los investigadores al momento de registrar los datos de su actividad científica o de sus identificadores en sus hojas de vida o perfiles de grupo. El código de los procesos de integración se puede encontrar en los recursos del Anexo B, siguiendo la ruta *scopus → "**integración.py**".* Este proceso también es utilizado por el *script* principal del proyecto.

Tras finalizar el proceso de integración, se extendieron las características de las tablas de autores y productos de *Scopus,* dado que sus registros vieron la necesidad de agregar nuevos datos correspondientes a los grupos de investigación provenientes de las bases de datos generadas en el módulo anterior, modificando ligeramente el diseño inicial de las dos tablas al agregar una columna nueva para "*idgruplac*" y otra para el nombre del grupo. De este modo los registros de autores y productos contienen a sus grupos de investigación asociados.

5.3.4. Interfaz gráfica

La interfaz gráfica desarrollada se puede observar en la Figura 5.6, cumpliendo funciones sencillas de extracción con respecto a códigos "*eid",* para los productos, y "*autor id",* para los autores de *Scopus.* La aplicación *web* cuenta con un apartado de instrucciones y requiere del ingreso de las credenciales "*API Key"* y "*API Institutional Token".* Los datos de los valores referenciados son persistidos en la base de datos "*scopusdb*" y actualizan la información de manera similar a la interfaz gráfica del Módulo Extractor *CVLAC-GrupLAC*. La aplicación notifica al usuario sobre el estado de la extracción y el manejo de excepciones mediante mensajes en la vista principal.

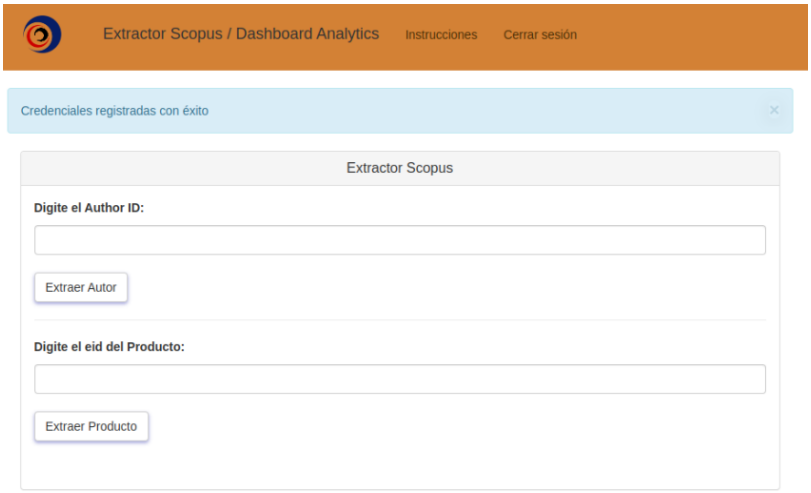


Figura 5.6. Interfaz gráfica de usuario del Módulo Extractor *Scopus*. Fuente propia.

La construcción de esta herramienta fue pensada para el contexto de la presente investigación, sin embargo, es capaz de extraer cualquier producto o autor con sólo ingresar su código respectivo. Un usuario administrador es capaz de modificar la extracción masiva de datos cambiando la lista de afiliaciones. Por defecto este conjunto de afiliaciones corresponde al generado para el Departamento del Cauca.

5.4. FASE DE EVALUACIÓN

5.4.1. Verificación

El presente módulo, a diferencia del módulo anterior, fue dotado de interfaces robustas y confiables como lo son las *APIs* de *Scopus* para la extracción de los datos bibliográficos de la actividad científica en el Cauca. El funcionamiento de los

métodos para el manejo de la clase extractora consiste únicamente en capturar y direccionar los datos que son ofrecidos por estas interfaces haciendo uso de llaves y valores específicos, prediseñados y documentados en las vistas de cada *API* [28]. La información trabajada se recuperó en forma organizada o estructurada y con una certeza absoluta en cuanto a la fiabilidad e integridad de los datos, puesto que estos provienen directamente de la base de datos de *Scopus.*

Los procesos de filtrado de datos o segmentación funcionan bajo las propiedades internas de *Scopus,* por lo que estos son ocultos. No obstante, ofrecen la seguridad y el nivel de sofisticación original de las plataformas de *Elsevier* que verifican su funcionamiento bajo la dirección de desarrolladores, diseñadores, evaluadores y arquitectos expertos [18], [69], [70]. En este orden de ideas, el funcionamiento interno del Módulo Extractor *Scopus* delegó las tareas más críticas a estas herramientas y dispuso sus partes originales a manera de vehículos de los datos para su selección y persistencia, sin interferir en su disposición o integridad original. Finalmente se asumió que la información extraída es lo suficientemente confiable y fiel a la fuente por tratarse de una herramienta propia de la firma, y que los mecanismos adoptados para su recuperación son aceptables dado que han superado una verificación exhaustiva manejada por equipos internos altamente capacitados.

5.4.2. Validación

La validación de los requisitos se desarrolló del mismo modo que en la sección 4.4.2, solicitando nuevamente la asistencia de los mismos investigadores. En la Tabla 5.5 se observa el resultado del proceso de validación.

Requisito	Primera Validación	Segunda Validación	Observación
1	✓	✓	
2	✓	✓	
3	✓	✓	
4	✓	✓	
5	✓	✓	
6	✓	✓	
7	✓	✓	
8	✓	✓	Considerar unificar bases de datos.

Tabla 5.1. Validación de requisitos del Módulo Extractor *Scopus*. Fuente propia.

6. MÓDULO *DASHBOARD*

6.1. FASE DE ANÁLISIS

La identificación de requisitos del Módulo *Dashboard* tomó como base las brechas encontradas en la revisión sistemática de la literatura. El estado actual del conocimiento permitió destacar diferentes herramientas de análisis bibliométrico de excelente calidad y fuentes prestigiosas de información bibliográfica reconocidas a nivel mundial. Del mismo modo, se consideraron herramientas y fuentes de menor reconocimiento, pero con un mayor grado de relación al contexto abordado en este trabajo de grado. A partir de la investigación se consideraron variedad de métodos y métricas para la construcción de una herramienta complementaria a las ya existentes, que aporte de manera sustancial al análisis bibliométrico de los grupos de investigación en el Departamento del Cauca.

Según lo expuesto en el planteamiento del problema, las herramientas disponibles de “La ciencia en cifras” otorgadas por Minciencias en [11], [12] poseen información desactualizada y no ofrecen la posibilidad de analizar detalladamente el desempeño de cada uno de los grupos de investigación en contraste con el panorama regional y nacional. Esta situación limita la extracción de conocimiento acerca de la actividad, participación, productividad, colaboración e impacto que tiene cada grupo, dificultando su análisis y una toma de decisiones provechosa para potenciar sus resultados. En las figuras 6.1 y 6.2 se observan ejemplos del análisis ofrecido por las herramientas mencionadas para el Departamento del Cauca, en donde se encuentran la caracterización y producción de los grupos de investigación al máximo nivel de granularidad permitido y con la fecha más reciente disponible. Cabe mencionar que la exploración de la información bibliográfica registrada en plataformas como *CVLAC* y *GrupLAC* es una tarea complicada dada la forzosa búsqueda manual para acceder a estos datos, puesto que no existía una herramienta para tal fin y los datos de los perfiles de investigación se presentan de manera individual y no estructurada.

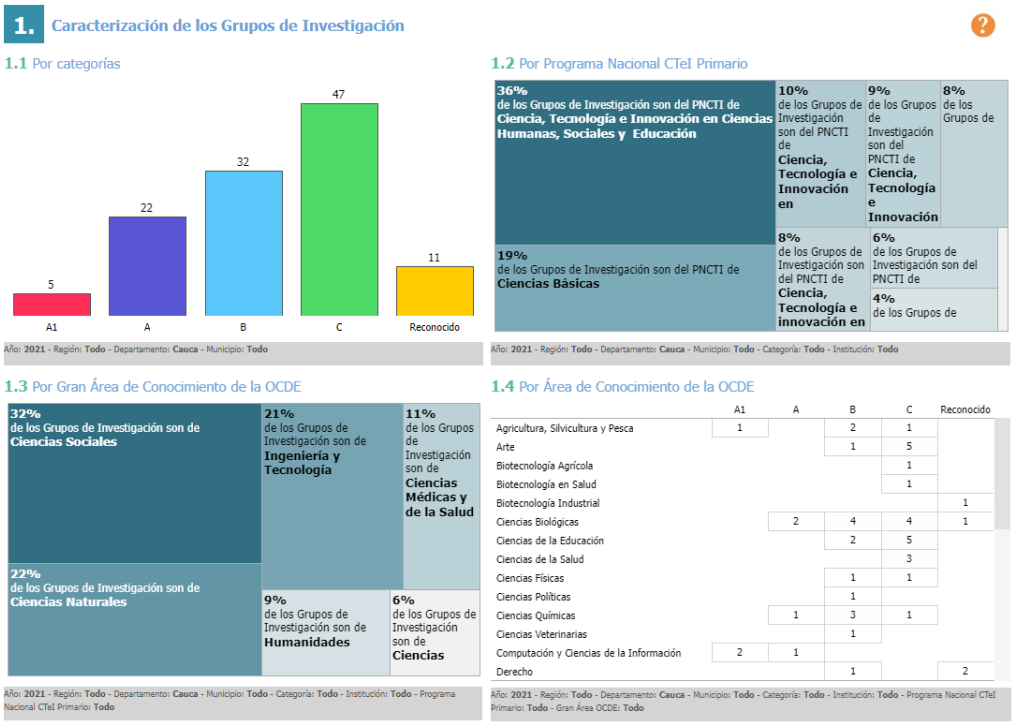


Figura 6.1. Caracterización de grupos de investigación en el Cauca según “La ciencia en cifras”. Tomado de [12]

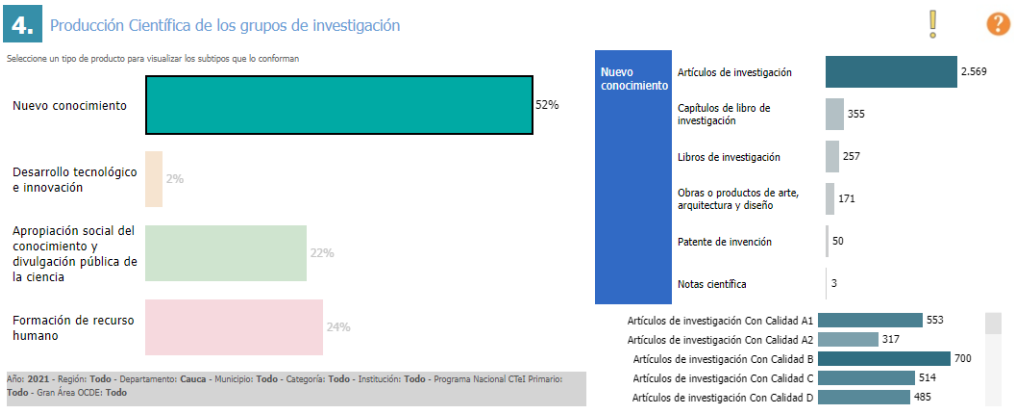


Figura 6.2. Producción de grupos de investigación en el Cauca según “La ciencia en cifras”. Tomado de [12]

Durante esta fase se analizaron las características necesarias para la construcción de un subsistema capaz de realizar la exploración, análisis y visualización

bibliométrica de los perfiles de los grupos y su producción científica en el Departamento del Cauca.

6.1.1. Selección de los datos

Las bases de datos generadas por los módulos explicados en los capítulos anteriores son explotadas por este módulo dado que contienen todos los conjuntos de datos seleccionados y extraídos para el contexto de la investigación científica a nivel regional. Las diferentes tablas de información bibliográfica que representan la producción científica, con sus múltiples dimensiones o características, se capturan en estructuras de datos que servirán como la entrada y materia del tablero de análisis. No obstante, no toda la información resultó apta para el análisis visual puesto que algunas características ofrecen una mayor relevancia que otras, además de encontrarse distintos tipos de datos entre las variables categóricas y numéricas de los productos de investigación. Es común que los conjuntos de datos contengan atributos irrelevantes, redundantes o desfavorables para el proceso de análisis, entendiendo a los atributos como las columnas de las tablas. Por este motivo, se tuvieron en cuenta los atributos necesarios y de mayor valor percibido en el estado actual del conocimiento a partir de las herramientas bibliométricas más destacadas y la literatura de la temática (Secciones 2.2.5 y 2.2.6). Para las secciones gráficas del tablero de análisis se seleccionaron los siguientes atributos:

- **Variables categóricas:** Tipo o Subtipo de Producto, País-Lugar, Revista, ISSN, Editorial, Institución-Afiliación, Líneas de Investigación, Clasificación de Grupo, Áreas de Conocimiento, Temática, Sexo, Sectores de Investigación, Palabras Clave de Autor, Palabras Clave Indizadas, Autores, Coautores, Plataforma-Ambiente.
- **Variables numéricas:** Citaciones, H1-index, Documentos.
- **Variables booleanas:** Producto Verificado, Aval, Certificación, Disponibilidad.
- **Otras variables:** Nombre-Título, Fecha, Identificador *GrupLAC*, Identificador *CVLAC*, Identificador *Scopus*.

Dado que la herramienta buscaba ofrecer, adicionalmente, un explorador de datos sobre las tablas en texto plano, se conservaron versiones completas de estas para tal fin. Cabe mencionar que los atributos escogidos para la construcción de las gráficas no están presentes en todas las tablas y fuentes de datos, sino sobre

aquellas que originalmente los poseen. Por ejemplo, no existen registros de citaciones sobre ninguno de los productos de *CVLAC* o *GrupLAC*, así como no todos los productos presentan variables categóricas como las palabras clave.

6.1.2. Selección de indicadores y métricas de análisis

El tablero de análisis cuenta con una sección de indicadores bibliométricos para la medición de diferentes aspectos de desempeño de los grupos de investigación tales cómo su productividad, actividad, impacto y participación basándose en su información bibliográfica. Estas métricas de análisis, comúnmente conocidas como *"Key Performance Indicators"* en el sector corporativo, ayudan a cuantificar el progreso hacia algún objetivo intencionado a lo largo del tiempo. Los indicadores presentados más adelante resultan de datos numéricos tomados de la producción y actividad científica registrada en el departamento sobre las fuentes consideradas con el fin de poder analizar distintas características de los grupos de investigación.

Estos indicadores fueron adaptados al contexto de grupos de investigadores a partir de las medidas bibliométricas aceptadas por la comunidad considerada en la literatura [21], [46], [47], [71], así como importantes iniciativas como "*Snowball Metrics*", utilizadas por *Elsevier* al ser aceptadas por expertos de todo el mundo y al demostrar su solidez y claridad en la definición de indicadores bibliométricos de la actividad científica [72], [73]. Adicionalmente, los indicadores pueden ser absolutos, al considerar todos los tipos de productos de un grupo de investigación, o relativos, al considerar sólo un tipo de producto especifico como artículos, libros, software, entre otros. El producto seleccionado en el cálculo de cada indicador depende de la petición del usuario. Los siguientes son los indicadores bibliométricos seleccionados y adaptados que fueron utilizados en la herramienta:

Consistencia

Mide la variación en el número de publicaciones de un grupo de investigación en cada año sucesivo dado su periodo de actividad, es decir, desde el año de su primera publicación hasta el más reciente. Un grupo es más "consistente" si sus publicaciones están casi igualmente distribuidas en lapsos anuales. Por ejemplo, un grupo tiene 40 publicaciones en el periodo entre 2015 y 2023. Si de las 40 publicaciones, en cada año se notificaron alrededor 4 a 6 publicaciones se dice que el grupo esta consistentemente publicando productos. Si, por otro lado, de las 40

publicaciones 32 fueron reportadas en 2020 y las restantes entre los otros años, entonces se dice que el grupo es poco consistente. La consistencia de un grupo de investigación está dada por:

$$\text{Consistencia} = \frac{1}{CVG}$$

Donde ***CVG*** es el "Coeficiente de Variación del Grupo" que permite obtener una medida de dispersión de los datos y está dado por:

$$\boldsymbol{CVG} = \frac{\sigma}{\overline{P}} \; ; \; \overline{P} \neq 0$$

Donde, $\boldsymbol{\sigma}$ = Desviación estándar de productos ; $\overline{\boldsymbol{P}}$ = Media aritmética de productos

$$\overline{\boldsymbol{P}} = \frac{\sum_{i=Y_s}^{Y_e} P_i}{t+1}$$

$$\boldsymbol{\sigma} = \sqrt{\frac{\sum_{i=Y_s}^{Y_e} {P_i}^2}{t+1} - \left(\frac{\sum_{i=Y_s}^{Y_e} P_i}{t+1}\right)^2}$$

Donde, $\boldsymbol{P_i}$ = Total de productos del grupo en el año *i* ; $\boldsymbol{t} = Y_e - Y_s =$ Periodo de actividad del grupo ; $\boldsymbol{Y_s} =$ Año inicial ; $\boldsymbol{Y_e} =$ Año final.

El valor de un $\boldsymbol{P_i}$ depende del tipo (o tipos) de producto a analizar por lo que puede significar, por ejemplo, el total de artículos de un grupo o el total de productos en general del grupo. Un valor cercano a 0 del *CVG* indica una menor dispersión de los datos del grupo. Para facilitar la lectura del indicador por parte del usuario final se decidió declarar la consistencia como inversamente proporcional al *CVG* de modo que una mayor dispersión resulta en una menor consistencia de la actividad del grupo. Las unidades del indicador son adimensionales, lo que resulta útil para comparar conjuntos de datos de poblaciones diferentes y facilitar la interpretación entre los datos de los diferentes grupos de investigación.

Promedio de Productos por Año:

El Promedio de Productos por Año (*PPA*) es un indicador que representa el número promedio de documentos o productos publicados dentro del periodo de tiempo de actividad de un grupo de investigación, ofreciendo una métrica de productividad de este. El indicador puede ser absoluto o relativo a un tipo de producto específico. El *PPA* está dado por:

$$\frac{\sum_{i=Y_s}^{Y_e} P_i}{(Y_e - Y_s) + 1}$$

Los nombres de estas variables se pueden observar en el apartado del indicador de Consistencia.

Porcentaje de Productos en los Últimos Años:

Indicador que representa el porcentaje de la producción científica de un grupo de investigación con respecto a la producción general de todos los grupos del departamento en los últimos años. Permite evidenciar la tendencia referente a la generación de nuevas producciones contrastando la participación de los grupos en el panorama regional. Se tienen en cuenta los últimos 3 años, no obstante, este valor se puede personalizar a cualquiera deseado. Como en anteriores casos, el indicador puede ser absoluto o relativo a un tipo de producto seleccionado. El Porcentaje de Productos en los Últimos Años (*PPUA*) está dado por:

$$\frac{\sum_{i=Y_s}^{Y_e} P_i}{\sum_{i=Y_s}^{Y_e} T_i} \times 100$$

Donde, $\boldsymbol{T_i}$ = Cantidad de productos de todos los grupos en el año *i*.

Conteo de Autores:

Mide la cantidad total de autores vinculados a un grupo de investigación. En el caso de la información proveniente de *GrupLAC* se trata de la cantidad de integrantes

visibles en el perfil. Por otro lado, la información capturada de *Scopus* permite identificar sólo los autores que pertenecen al grupo y que al mismo tiempo han registrado adecuadamente su identificador *Scopus* en la hoja de vida de *CVLAC*. El valor de este indicador no varía con el tipo de producto seleccionado.

Conteo de Productos:

Mide la cantidad total de productos generados por un grupo de investigación, sean estos todos los tipos de productos o un tipo de producto en específico. De manera similar al indicador anterior, esta medida puede representar valores diferentes entre los conjuntos de datos de *GrupLAC* y *Scopus* dependiendo de la cantidad de registros correctamente diligenciados (completos y acertados) en el perfil del grupo y por tanto de la cantidad de productos emparejados e identificados en *Scopus*.

Indicadores de Minciencias:

Los indicadores utilizados por Minciencias para la categorización de grupos de investigación [21], son referenciados para el grupo en cuestión y ofrecidos al usuario para su visualización en el navegador de internet a través de un enlace disponible en el *Dashboard.* Como sugiere la documentación, es posible observar indicadores absolutos de cohesión, cooperación, trayectoria, estabilidad, entre otros más. Estos indicadores son calculados directamente por Minciencias por lo que presentan el inconveniente de basarse, hasta la fecha, en “información registrada y actualizada a 20 de octubre 2021 (fecha de cierre de la Convocatoria 894 de 2021)”.

Citaciones por Producto:

Calcula el promedio de citaciones que tienen los productos emparejados a un grupo de investigación en *Scopus* hasta la fecha de la última extracción de datos realizada. La medida de las citaciones funciona como indicador del impacto y presencia que ha tenido el grupo de investigación mediante producciones publicadas por las revistas o editoriales de mayor reconocimiento internacional. Este indicador está dado por:

$$\frac{\sum_{j=1}^{NTP} C_j}{NTP}$$

Donde, $\boldsymbol{C_j} =$ Número total de citaciones del producto j ; $\boldsymbol{NTP} =$ Número total de productos del grupo.

Conteo de Citaciones:

Calcula la suma total de citaciones de todos los productos emparejados a un grupo de investigación en *Scopus* hasta la fecha de la última extracción de datos realizada.

Índice $\boldsymbol{h}$ de primer orden:

El índice h, originalmente propuesto por Hirsch en 2005 [74] permite medir la productividad y el impacto de un investigador basado en la relación entre sus productos y las citaciones de estos. Este índice está presente en casi todas las herramientas consideradas en el estado actual del conocimiento. Si un investigador ha generado h productos con al menos h citaciones cada uno, entonces su índice es igual a h. La siguiente gráfica permite observar el índice h a partir de la relación de los productos y sus citaciones para un autor.

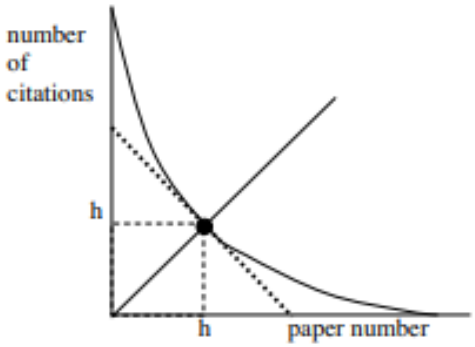


Figura 6.3. Representación gráfica del índice h. Tomado de [74]

Es posible adaptar el índice h a un grupo de investigación como sugiere [71] para los índices h_1 y h_2 o índice h de primer orden e índice h de segundo orden. El índice h_1 de un grupo de investigación es igual a h si el grupo ha publicado h productos emparejados en *Scopus,* cada uno teniendo al menos h citaciones. Por ejemplo, si un grupo tiene 3 productos (p_1, p_2, p_3) con citaciones $(2, 5, 10)$ entonces su índice h_1 es igual 2.

Índice h de segundo orden:

El índice h_2 de un grupo de investigación es igual a h si el grupo tiene h investigadores emparejados en *Scopus*, cada uno de ellos teniendo un índice h individual cuyo valor es de al menos h. Por ejemplo, si un grupo tiene 5 integrantes $(I_1, I_2, I_3, I_4, I_5)$ con índices h individuales de $(10, 5, 8, 6, 16)$ entonces su índice h_2 es igual a 5. Este puede ser considerado como uno de los indicadores bibliométricos de los grupos de investigación más estrictos entre los planteados.

6.1.3. Procesamiento de los datos

Los datos extraídos se manejaron a través de una etapa inicial de preprocesamiento que consiste en la limpieza y formateo general de estos. Las estructuras de datos de todas las fuentes utilizadas deben adaptarse de manera adecuada para su posterior exploración, análisis y visualización, por lo que su preparación incluye:

- Eliminación de duplicados con base en identificadores de autor y producto como *DOI*, *ISBN, AuthorID, ScopusID* o códigos de *GrupLAC*. También con base en combinaciones únicas de características tales cómo títulos, autores, fechas, editoriales o institución.
- Normalización de cadenas de texto de las diferentes características de los registros según acentos, mayúsculas, símbolos, espacios, saltos de línea, etiquetas, caracteres de escape, etc.
- Manejo de valores atípicos.
- Manejo de valores nulos.
- Formateo de los tipos de datos como fechas, booleanos, flotantes, cadenas de texto, etc.
- Definición de diccionarios para la unificación de nombres de entidades como revistas, editoriales, afiliaciones, instituciones, agencias, entre otras.

La limpieza de los datos fue más exhaustiva para las fuentes de *ScienTI* puesto que el registro de los datos es mucho menos restrictivo y cuidadoso para el etiquetado y manejo de excepciones que se pueden presentar en el ingreso de la información por parte de los investigadores. Como resultado, los datos extraídos de estas plataformas presentaron una mayor cantidad de inconsistencias y errores a manejar en comparación con los recuperados de la base de datos de *Scopus.* Estos últimos

presentaron inconvenientes en las variables categóricas como editoriales, revistas e instituciones, y evidencias de multiplicidad de perfiles de investigador para el mismo autor.

El procesamiento de datos subsecuente comprende el filtrado de la información a partir de la fuente de datos, elementos o productos de interés, y características particulares de estos. De este modo se ofrecieron varios niveles de detalle cuyos valores son determinados por el usuario para la generación del análisis. Los grupos de investigación pueden ser analizados de manera individual o en conjunto sobre su producción científica en general o sobre algún tipo de producto especifico. El análisis incluye la temporalidad, proporcionalidad, distribución, contraste y emparejamiento de los datos de modo que sea posible extraer e interpretar conocimiento acerca de la evolución, colaboración, impacto, actividad, participación, productividad y otras cualidades de los grupos de investigación con base en sus registros. Adicionalmente, se cuenta con un motor de cálculo para los indicadores anteriormente mencionados, que toma la solicitud del usuario para definir valores absolutos o relativos de las métricas para uno o varios grupos según sea el caso, ofreciendo así también la posibilidad de generar reportes comparativos.

6.1.4. Visualización de los datos

La definición de los tipos de gráficas a utilizar va ligada a los tipos de datos en cuestión y al resultado esperado en el análisis. Siguiendo las recomendaciones observadas en el estado actual del conocimiento sobre el análisis bibliométrico y considerando los datos disponibles se planteó el uso de series de tiempo sobre registros que cuentan con valores de fecha. Para los diferentes atributos que definen categorías en proporciones o frecuencia se utilizaron gráficas de tipo pastel y diagramas de barras. Se utilizaron “*treemaps*” para visualizar datos jerárquicos según la participación de elementos como autores o temas trabajados en un contexto dado. La colaboración de los grupos de investigación se observa mediante un gráfico de red de conexiones entre nodos con enlaces de peso variable. La distribución de citaciones a través de sus diferentes cuartiles o según los grupos de investigación y los tipos de producto, se representaron mediante diagramas de cajas y mapas de calor que pueden contener uno o varios grupos a la vez.

Los diferentes indicadores bibliométricos son observados mediante cajas de información junto con enlaces de interés a los perfiles de los grupos y a los

indicadores oficiales de Miniciencias. Para estos indicadores se destinaron igualmente gráficas comparativas a modo de diagrama de barras y gráficos de radar. Las gráficas definidas son interactivas, de modo que cuentan con funciones como "*zoom*", auto escalado, desplazamiento, descarga en formato de imagen, ocultar o mostrar elementos, ampliar la información al sobreponer el cursor y transiciones animadas.

Las gráficas se construyeron a partir de la fuente y valores especificados por el usuario para lo cual se destinan secciones de filtrado de datos con espacios de información ordenados de mayor a menor granularidad. Los grupos de investigación identificados en *GrupLAC* y *Scopus* pueden ser filtrados individualmente o en conjunto, de modo que se pueden ingresar de manera manual o a través de atributos como su institución, clasificación, áreas y líneas de investigación. También es posible seleccionar el tipo de producto de interés o todos los productos en general.

El explorador de datos incluye todas las fuentes, productos y características seleccionadas. Este recupera y presenta la información a manera de tablas según la solicitud definida en el filtro. Los valores especificados en el último nivel de detalle dependen de la característica seleccionada por lo que pueden ser rangos de fecha, etiquetas, cadenas de texto o valores numéricos. Por ejemplo, si se selecciona "Artículos" como elemento y "Revista" como característica, el ultimo nivel de detalle permitirá seleccionar una o varias etiquetas correspondientes a cada uno de los valores únicos de revistas disponibles en los registros. Por otro lado, si se selecciona "Empresa Tecnológica" y "Tipo", esta combinación dará lugar a etiquetas como "*Start-up*" o "*Spin-Off*". Otras características como "Título" o "Fechas" reciben cadenas de texto o rangos respectivamente. Sobre este orden de ideas, son muchas las combinaciones posibles y entradas que puede otorgar un usuario para explorar eficientemente los datos registrados a nivel regional entre las fuentes de *CVLAC*, *GrupLAC* y *Scopus*.

6.1.5. Definición de requisitos

Con base en lo considerado durante la fase de análisis se definieron los siguientes requisitos para el subsistema del módulo *Dashboard:*

- Preprocesar los datos provenientes de las bases de datos generadas por los módulos extractores.
- Filtrar la información a través de varios niveles de detalle.
- Explorar los datos disponibles a lo largo de todos los conjuntos de datos generados por los módulos extractores.
- Calcular y presentar los indicadores bibliométricos seleccionados.
- Generar estadísticas y gráficas interactivas para el análisis y visualización bibliométrica de los grupos de investigación en el Cauca con la información disponible.
- Hacer un correcto manejo de excepciones y notificaciones durante el análisis.

6.2. FASE DE DISEÑO

6.2.1. Flujo de trabajo

La secuencia de procesos de principio a fin para el uso del *Dashboard,* se resume mediante el flujo de trabajo definido en la Figura 6.4.

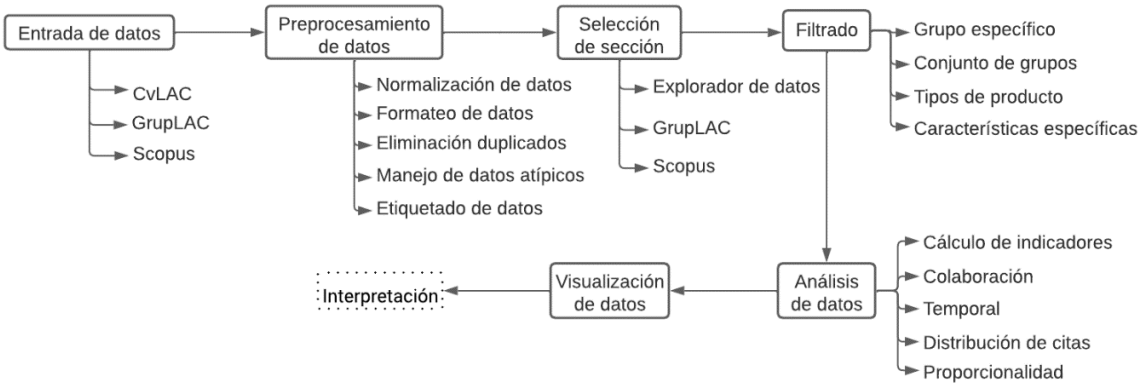


Figura 6.4. Flujo de trabajo del *Dashboard*. Fuente propia.

Como se observa, el flujo de trabajo se encuentra dividido en 7 etapas principales:

1. ***Entrada de datos:*** Los datos extraídos desde las fuentes de *CVLAC*, *GrupLAC* y *Scopus* son recuperados directamente desde el conjunto de bases de datos generado. Se utilizan estructuras de datos de la librería *Pandas* y en formato *CSV* a manera de copias de las versiones más actualizadas de las tablas de información seleccionadas y persistidas.

2. ***Preprocesamiento de datos:*** Esta etapa toma la información recuperada y la encamina por los procesos de limpieza mencionados en la sección 6.1.3. Posteriormente se obtienen nuevos conjuntos de tablas de información adecuados y listos para los procesos de análisis y visualización. Esta etapa incluye la actualización de la información a la versión más reciente de las bases de datos si se detectan cambios.

3. ***Selección de sección:*** El usuario tiene disponible 4 secciones diferentes:

   a) ***Home:*** Es la sección informativa del *Dashboard* en donde se presentan el contexto y los autores del proyecto.
   b) ***Explorador de datos:*** Sección que dispone la herramienta exploradora de datos de las fuentes bibliográficas. En ella se presenta una subsección de filtrado y otra de presentación de los datos. El filtro del explorador de datos permite seleccionar la fuente de los datos, el elemento o producto específico de interés, la característica especifica por la cual se desean filtrar registros, y el valor particular a digitar para dicha característica, o bien las etiquetas si se trata de valores categóricos. Una vez filtrados los datos, las coincidencias se muestran mediante tablas en el espacio de presentación ofreciendo la información de todos los registros y columnas disponibles.
   c) ***GrupLAC:*** Sección del *Dashboard* que permite realizar el análisis y visualización bibliométrica de los grupos de investigación de manera individual o de manera general, tomando como referencia los datos extraídos y preprocesados para las tablas de información de los perfiles de investigación en *GrupLAC*. Cuenta con su panel de filtrado en donde se puede seleccionar el tipo de análisis, el grupo (o grupos) de investigación y el elemento (o elementos) de interés. Presenta los indicadores bibliométricos seleccionados para esta fuente de datos, enlaces de internet del grupo y variedad de gráficas interactivas construidas a partir de los valores especificados por el usuario.
   d) ***Scopus:*** Esta sección es similar a la sección de *GrupLAC*, con la diferencia de que los datos se toman a partir de la integración realizada sobre *Scopus* que luego es preprocesada. Los grupos de investigación presentados en esta sección son aquellos que lograron ser identificados en la base de datos bibliográfica *Scopus* a partir de los procesos planteados en anteriores capítulos, tratándose de la información

bibliográfica de mayor reconocimiento con presencia en revistas y editoriales de alto impacto. El panel de filtrado también es similar al de la sección anterior, no obstante, se tienen diferencias en los indicadores y gráficas presentadas dado que se aprovecha la información extra que ofrece *Scopus* como números de citaciones e índices *h* de autores.

4. ***Filtrado:*** Las secciones de exploración y análisis de datos contienen sus propios paneles de filtrado con diferentes campos de información que corresponden a los niveles de detalle. Esta etapa consiste en la tarea de ingresar los valores de dichos campos para segmentar los datos con los que el *Dashboard* realiza su trabajo. Los campos corresponden a diferentes parámetros como fuentes de datos, tipos de producto, características específicas, grupos de investigación, etiquetas, rangos de fecha, entre otros.

5. ***Análisis de datos:*** Etapa encargada del cálculo de indicadores bibliométricos y del análisis de datos cualitativo y cuantitativo de la información segmentada que puede ser referente a variables de tiempo, frecuencia, etiqueta, colaboración, impacto, productividad, etc. El análisis se puede dar de manera individual o en conjunto para los grupos de investigación.

6. ***Visualización de datos:*** Etapa encargada de la representación gráfica del análisis realizado. Las gráficas brindan información de: indicadores bibliométricos, productos anuales generados, publicaciones en revistas o editoriales, publicaciones por tipo, participación de autores, red de colaboración, distribución de citaciones, temas trabajados y comparativos de estos y más elementos entre los grupos de investigación. Se presentan variaciones de las gráficas mencionadas y gráficas adicionales según la información ingresada en los filtros y las características disponibles en las estructuras de datos.

7. ***Interpretación:*** Consiste en la extracción de conocimiento por parte del usuario.

6.2.2. Estructura y funcionalidad

La interfaz de usuario del *Dashboard* consta de diferentes elementos estratégicamente dispuestos para optimizar la visualización y manipulación de los datos. La estructura general se muestra en las figuras 6.5 y 6.6.

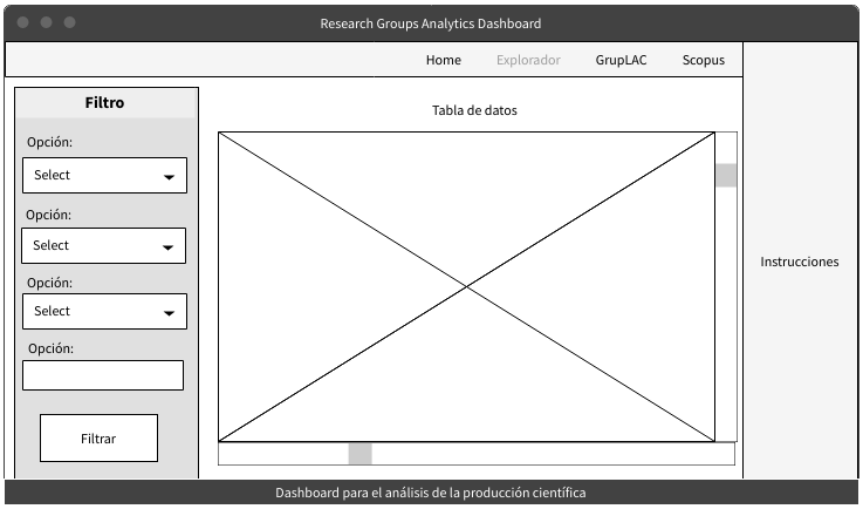


Figura 6.5. *Wireframe* para la sección explorador. Fuente propia.

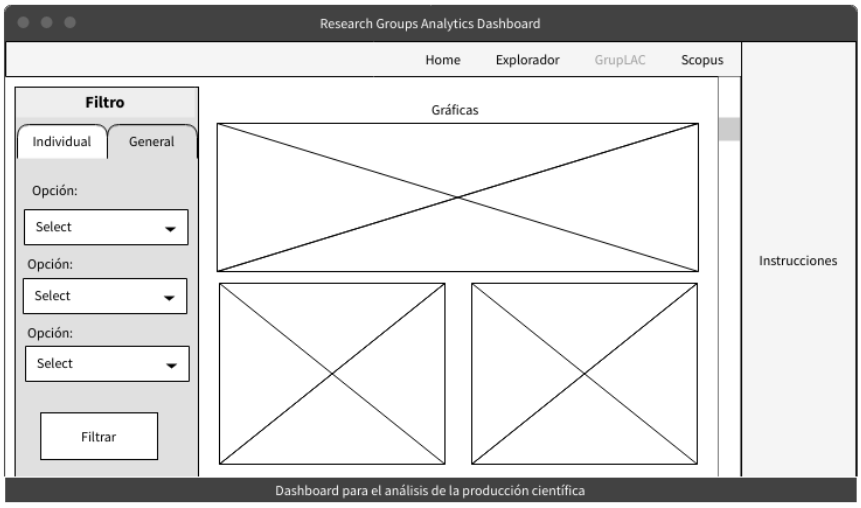


Figura 6.6. *Wireframe* para las secciones *GrupLAC y Scopus*. Fuente propia.

El *Dashboard* cuenta con pestañas de navegación ubicadas siempre visibles en la parte superior de la página web, un apartado de visualización para cada sección de

análisis de los datos ubicado en el área central, un panel de filtrado en la parte izquierda y un apartado de instrucciones mostrado como barra lateral desplegable en el extremo derecho, accesible desde cualquier sección de la plataforma. El usuario puede interactuar con los diferentes elementos de los apartados, ingresar información y desplazarse por el sitio como en cualquier página *web* para llevar a cabo las tareas particulares descritas en el flujo de trabajo. El análisis y visualización bibliométrica se genera de manera automática una vez se ejecuta el filtrado.

6.2.3. Diseño de despliegue

A continuación, se presenta el diseño definido para el despliegue del subsistema que permite profundizar en la interacción de los componentes encargados de realizar las tareas principales del flujo de trabajo a petición del usuario.

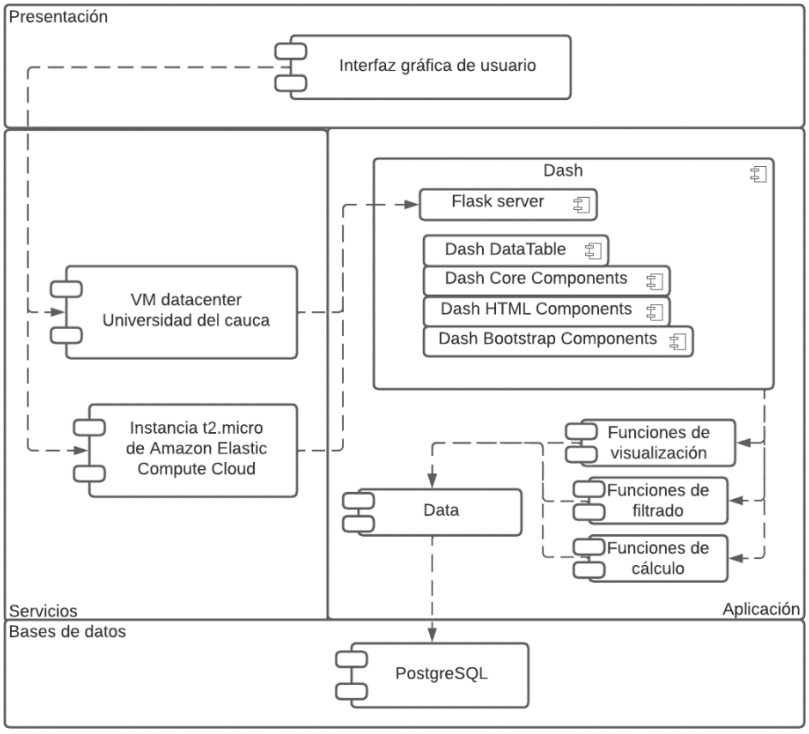


Figura 6.7. Diagrama de despliegue del *Dashboard*. Fuente propia.

El diagrama de la Figura 6.7 presenta diferentes componentes contenidos en 4 niveles:

1. ***Nivel de Presentación:***

    - ***Interfaz Gráfica de Usuario (GUI):*** Presenta la aplicación e interactúa con el usuario a través de un navegador web.

2. ***Nivel de servicios:***

    - ***Instancia t2.micro de Amazon Elastic Compute Cloud (Amazon EC2):*** Proporciona capacidad de computación escalable bajo demanda en la nube de *Amazon Web Services (AWS)* [75].
    - ***VM Datacenter red Telco:*** Máquina virtual disponible dentro de la red local "*Telco"* perteneciente a las instalaciones de la Universidad del Cauca.

3. ***Nivel de Aplicación:***

    - ***Dash:*** *Framework* de *Python* para crear aplicaciones *web* especializadas en análisis de datos [76].

        - ***Flask Server:*** Servidor *web* de *Flask* utilizado por *Dash* para correr el *Dashboard* en un navegador web.
        - ***Dash DataTable:*** Componente propio de *Dash* para mostrar datos en forma de tablas interactivas.
        - ***Dash Core Component:*** Componente principal de *Dash* para crear elementos interactivos y reactivos.
        - ***Dash Bootstrap Component:*** Componente propio de *Dash* para integrar Bootstrap en la aplicación *Dash.*
        - ***Dash HTML Component:*** Componente proprio de *Dash* para integrar HTML personalizado en la aplicación *Dash*.

    - ***Funciones de Filtrado, Cálculo y Visualización:*** Funciones específicas de la aplicación para procesar y presentar los datos.
    - ***Data:*** Componente que contiene y prepara los datos utilizados por la aplicación, se conecta con los componentes para el filtrado, cálculo y visualización de datos.

4. ***Nivel de bases de datos:***

    - ***PostgreSQL:*** Sistema de gestión de las bases de datos.

6.3. FASE DE CODIFICACIÓN

El código desarrollado para este subsistema se puede encontrar en el repositorio compartido en el Anexo B, dentro del directorio "*dashboard*".

6.3.1. Entrada y Preprocesamiento de datos

La implementación de los procesos de limpieza mencionados en la sección 6.1.3, se llevó a cabo mediante el uso de las librerías *pandas*, *numpy*, *json* y *regex* de *Python*. Se hizo uso de expresiones regulares personalizadas, codificadores y decodificadores de texto sobre estándares "*ASCII*" y "*UTF-8*", múltiples funciones sobre cadenas de texto de *Pandas*, diccionarios personalizados, técnicas de agrupación de datos, formatos de datos de *Pandas,* técnicas de emparejamiento datos y técnicas de detección de valores inconsistentes para fechas y datos numéricos. Las técnicas mencionadas se adaptaron puntualmente sobre cada reto particular que presentaron las múltiples combinaciones de tablas de información seleccionadas y cada una de sus dimensiones. En el Algoritmo 10 del Anexo C se encuentra el pseudocódigo del preprocesamiento de datos desarrollado a manera de resumen teniendo en cuenta la gran diversidad de tareas necesarias. Para la entrada de los datos se destinó el directorio "*assets*" → "***data***" en donde se almacenan los datos recuperados de las bases de datos y posteriormente los datos preprocesados. El preprocesamiento de los datos puede ser ejecutado utilizando el *script* "***preprocessing.py***" como se indica en el repositorio del proyecto en el Anexo B, conteniendo este todos los procesos codificados de la sección mencionada al inicio de este apartado.

6.3.2. Filtrado de datos

Los paneles de filtrado desarrollados para las secciones del *Dashboard* presentan los niveles de granularidad disponibles para la segmentación de datos según los intereses del usuario. Los campos de información responden de manera reactiva actualizando la información que se puede seleccionar o digitar. En la Figura 6.8 se observan 4 imágenes de los filtros, siendo **(*a*)** y **(*b*)** correspondientes al filtro del explorador de datos. La imagen en **(*c*)** corresponde al filtro de la sección *GrupLAC* y **(*d*)** al filtro de la sección *Scopus*, estos dos filtros son muy similares puesto que presentan diferencias únicamente en la información opcional para el usuario por ser

diferentes conjuntos de datos. Igualmente, se observan las opciones “Individual” y “General” para elegir una preferencia en el tipo de análisis que se llevará a cabo.

Figura 6.8. Paneles de filtrado de las secciones del *Dashboard*. Fuente propia.

Los filtros de *GrupLAC* y *Scopus* se enfocan en obtener la información referente a uno o varios grupos de investigación para realizar el análisis y visualización bibliométrica, mientras que el filtro del explorador de datos busca presentar registros específicos disponibles entre todos los conjuntos de datos extraídos.

6.3.3. Análisis y visualización de datos

Las funciones de cálculo de los indicadores bibliométricos y del análisis de los diferentes productos, autores y sus características, se desarrollaron con ayuda de la librería *Pandas* para la manipulación de estructuras de datos y el uso de herramientas de cálculo estadístico. En las imágenes de la Figura 6.9 se presentan los indicadores absolutos obtenidos para un grupo de investigación, siendo ***(a)*** la sección de *GrupLAC*, ***(b)*** la sección de *Scopus* y ***(c)*** un fragmento de los resultados oficiales de indicadores de Minciencias del grupo a su fecha más actual manejada.

(a)

(b)

(c)

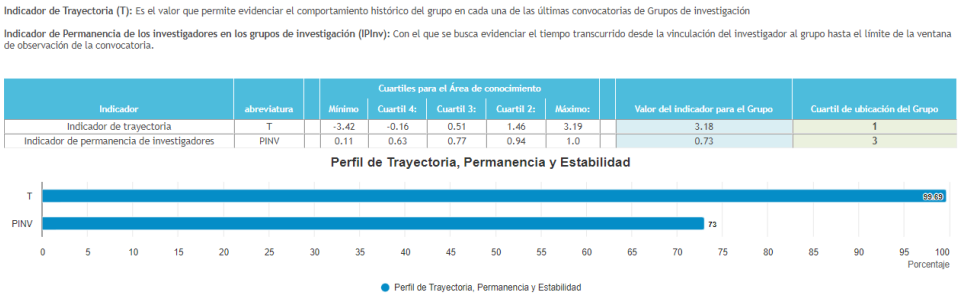
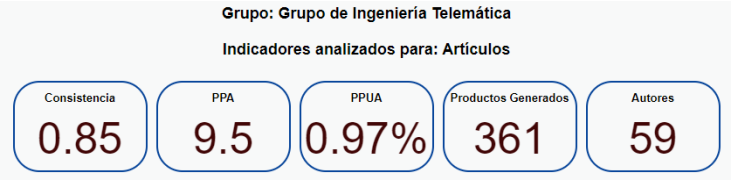


Figura 6.9. Indicadores absolutos obtenidos para el Grupo de Ingeniería Telemática en secciones *GrupLAC* **(a)**, *Scopus* **(b)** y en Minciencias **(c)**. Fuente propia para **(a)** y **(b)**, **(c)** es tomado de [77]

Cabe mencionar que los indicadores relativos de cada tipo de producto se obtienen filtrando dicho elemento. En la Figura 6.10 se presenta un ejemplo con los artículos del mismo grupo de investigación usado en los ejemplos anteriores, notándose las diferencias en algunos de los indicadores.

(a)



(b)

Figura 6.10. Indicadores relativos obtenidos para el Grupo de Ingeniería Telemática en las secciones *GrupLAC* **(a)** *y Scopus* **(b)**. Fuente propia.

Una vez se han filtrado y analizado los datos internamente en la herramienta por acción de un usuario, se procede a presentar los resultados mediante múltiples gráficas tras una pequeña espera. La Figura 6.11 muestra un ejemplo de la tabla de información generada para el filtrado de la Figura 6.8-a.

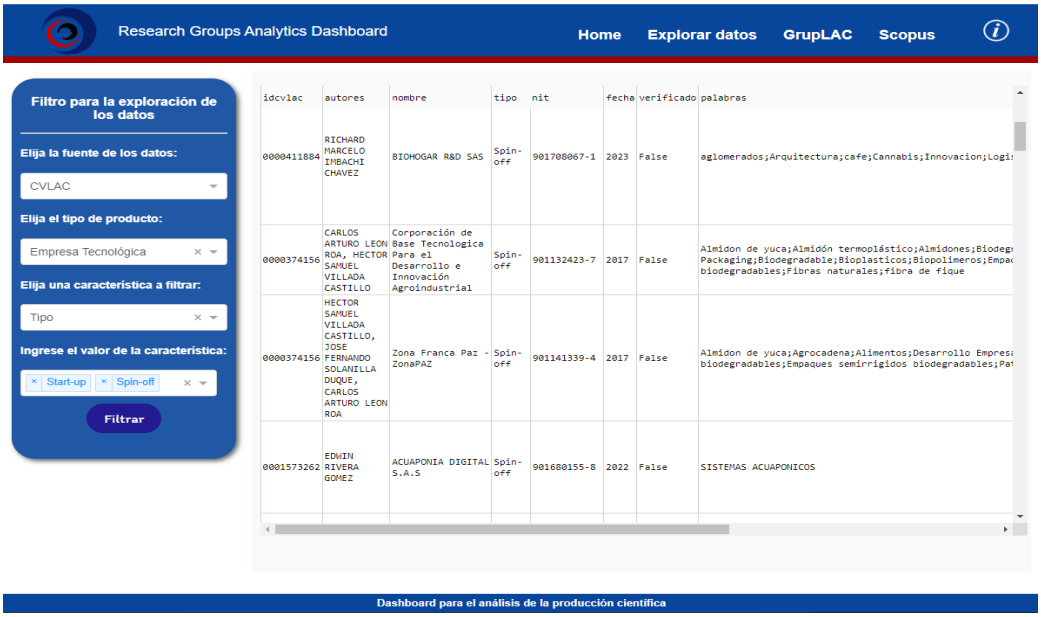


Figura 6.11. Ejemplo del uso del explorador de datos: Tabla de información. Fuente propia.

Las figuras 6.12 y 6.13 muestran fragmentos del análisis generado para los filtrados de las figuras 6.8-c y 6.8-d respectivamente. En el Anexo E se muestran más

ejemplos de gráficas generadas como red de colaboración, gráfico de barras, diagrama de cajas y *treemap* para otros valores de filtrado.

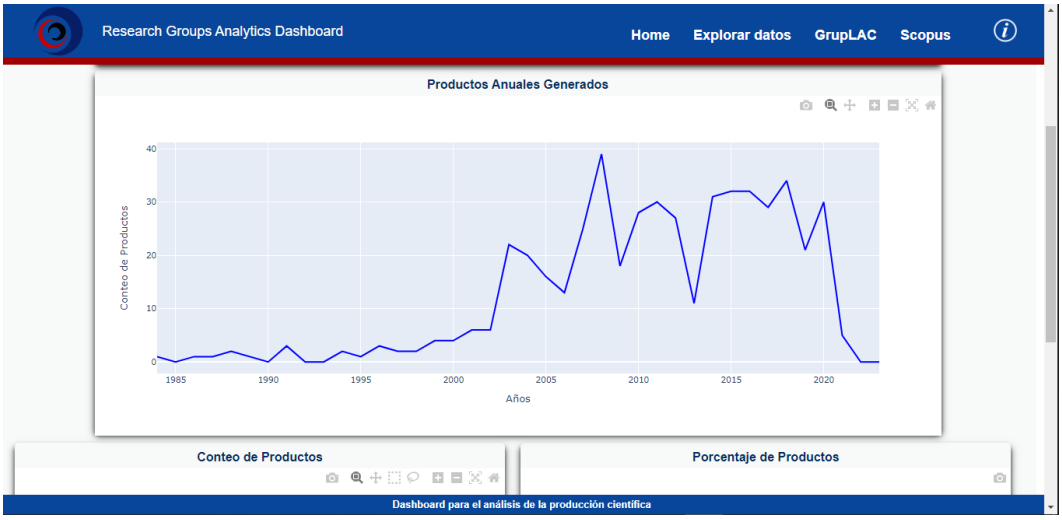


Figura 6.12. Ejemplo del análisis generado por la sección *GrupLAC* del *Dashboard*: Serie de tiempo. Fuente propia.

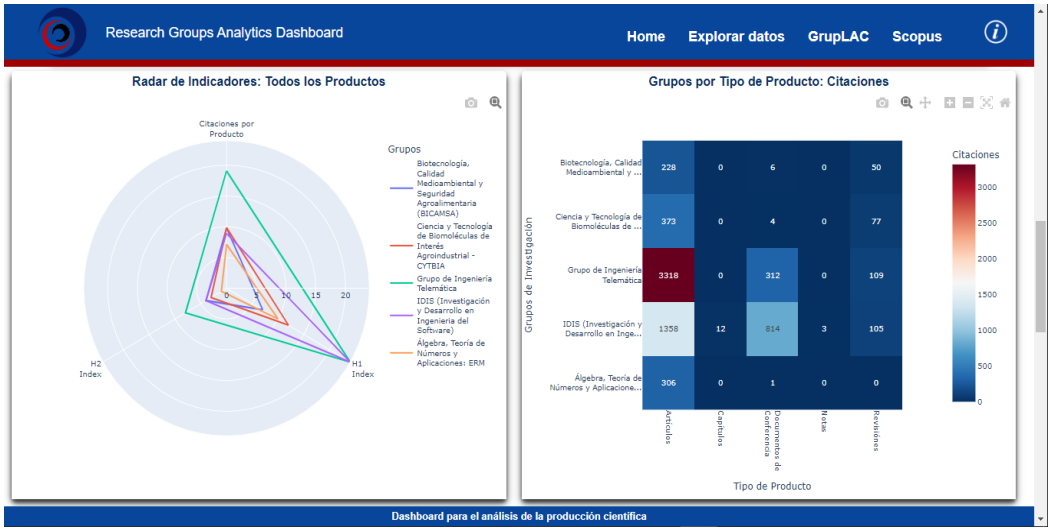


Figura 6.13. Ejemplo del análisis generado por la sección *Scopus* del *Dashboard*: Radar y mapa de calor. Fuente propia.

Las gráficas son interactivas y también permiten conocer detalles estadísticos o información puntual al dirigir el cursor sobre los componentes de estas. Los ejemplos no cubren el total de gráficas que pueden generarse puesto que estas dependen de los valores y combinaciones de los filtros. Dado que algunos

resultados pueden incluir muchos grupos de investigación, las gráficas aceptan un máximo de 10 grupos a la vez para evitar saturarse y corromper el análisis visual.

6.4. FASE DE EVALUACIÓN

El proceso de evaluación del *Dashboard* se describe a lo largo del Capítulo 7 de la Monografía. Dado que este módulo representa la última iteración de la metodología aplicada, la última fase contempla el incremento final e integración con el resto de los subsistemas. En el Capítulo 7 también se describe el proceso de despliegue del sistema y su adaptación para pruebas experimentales con el fin de revisar el cumplimiento de los requisitos del *Dashboard* y analizar los resultados.

7. PROTOTIPO Y EXPERIMENTACIÓN

7.1. INTEGRACIÓN DE SUBSISTEMAS

Los módulos desarrollados, como conjunto, poseen las propiedades de bajo acoplamiento y alta cohesión al tener un nivel bajo de dependencia entre los subsistemas y un alto grado de relación entre estos para los objetivos de la herramienta. De esta manera se puede configurar el flujo de datos de manera sencilla y permitir el funcionamiento general del prototipo, aunque alguno de los módulos llegara a fallar. En la Figura 7.1 se presenta la integración de los módulos teniendo en cuenta las estructuras internas de cada uno de ellos, presentadas en las fases de diseño de los capítulos 4, 5 y 6.

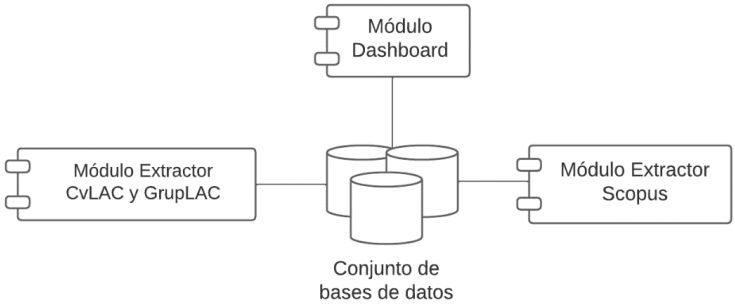


Figura 7.1. Diagrama de componentes de integración del sistema. Fuente propia.

Como resultado se obtuvo un prototipo funcional y escalable del *Dashboard* que cuenta con métodos de soporte para la actualización de datos a través de los extractores. Sobre el objetivo de la evaluación, se pone el foco en el módulo de análisis y visualización bibliométrica.

7.2. DESPLIEGUE DEL SISTEMA

7.2.1. Máquina Virtual

Se realizó la solicitud de una máquina virtual al administrador del *Datacenter* perteneciente al Departamento de Telemática de la Universidad del Cauca. La máquina virtual fue creada y asignada 4 días después de presentar la solicitud del recurso. Posteriormente, se accedió a la máquina virtual utilizando el *software* "*Anydesk"*, el cual permite el acceso remoto con las credenciales correspondientes

[78]. A través de este *software*, se llevó a cabo la instalación del proyecto y se procedió con la configuración para asignar una dirección *IP* y un puerto en el que se levantó el servidor.

Debido a las limitaciones de los recursos proporcionados por el *Datacenter*, el acceso al *Dashboard* quedó restringido únicamente a la red local "*Telco"* de la Facultad de Ingeniería Electrónica y Telecomunicaciones. Para superar esta limitación y permitir a los usuarios acceder al *Dashboard*, se solicitó autorización para utilizar las salas de cómputo que contaban con puntos de acceso *Ethernet* de dicha red. Para extender el acceso a través de la red local, se instaló un *router* que brindó una conexión inalámbrica para cualquier equipo dentro de la cobertura. Con esto listo, fue posible utilizar la herramienta desde varios ordenadores de las salas de cómputo simultáneamente y con un desempeño optimo.

7.2.2. Instancia EC2

Utilizando una cuenta de *Amazon Web Services* (*AWS*), se implementó un servicio de *Amazon EC2 (Amazon Elastic Compute Cloud)*. El tipo de instancia utilizado fue "*t2.micro"*, el cual forma parte de la capa gratuita ofrecida por *AWS* [75], [79]. Se accedió a la instancia a través del protocolo *Secure Shell (SSH)* para la instalación del *Dashboard* y su configuración para el acceso desde cualquier dispositivo con conexión a internet a través de una *IP* pública. Con la capa gratuita se establecen ciertos límites en los recursos disponibles para una instancia "*t2.micro"* en *Amazon EC2*. Estos límites incluyen 1GB de memoria RAM, 1GB de almacenamiento y un máximo de 750 horas de uso (las horas cuentan y se suman desde cada acceso). Dadas estas restricciones, se estableció destinar esta instancia para evaluaciones con un máximo de 2 usuarios simultáneos. Esta configuración cuenta con la ventaja de permitir el acceso al *Dashboard* desde cualquier dispositivo con conexión a internet.

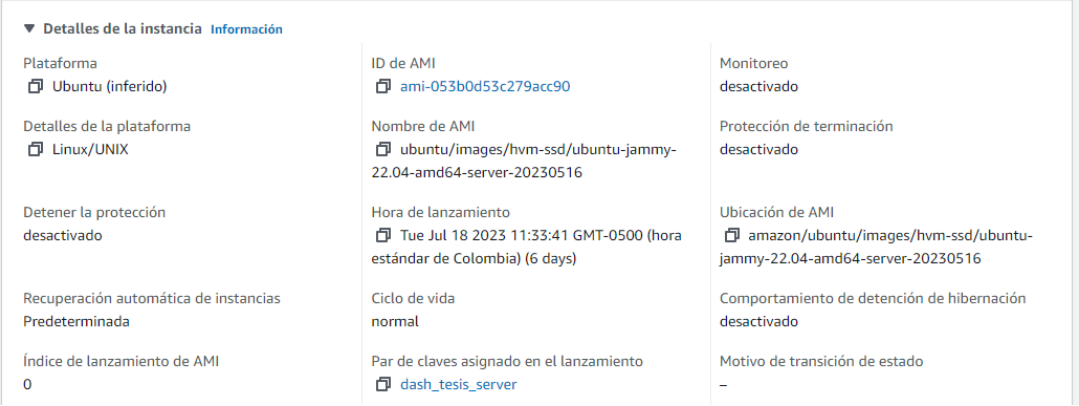


Figura 7.2. Detalles de la instancia *EC2* implementada. Fuente propia.

7.3. EXPERIMENTACIÓN

La evaluación del prototipo propuesto siguió una adaptación de la metodología de Experimentación en Ingeniería de Software [80] que comprende una serie de pasos para determinar la satisfacción de los interesados al interactuar con el sistema y su usabilidad. Los pasos del proceso se observan en la Figura 7.3. Estos consistieron en definir el objetivo de la experimentación, determinar el ambiente de ejecución del experimento, elegir a las personas que realizaron la evaluación, establecer los instrumentos necesarios, ejecutar las pruebas y finalmente llevar a cabo el análisis e interpretación de los resultados. Los pasos se profundizan a lo largo de esta sección, con excepción del último paso expuesto en la sección 7.4.

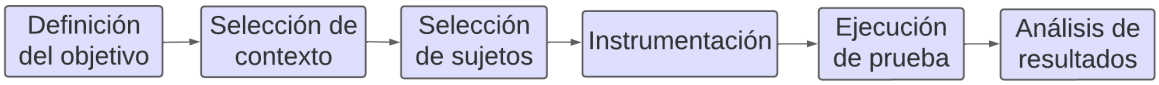


Figura 7.3. Diagrama del proceso de experimentación. Fuente propia

7.3.1. Definición del objetivo

Con el fin de asegurar el cumplimiento de los requisitos de la herramienta y realizar una correcta evaluación, se definió como objetivo de la experimentación: analizar las características del *Dashboard* con el fin de evaluarlas respecto a la usabilidad y satisfacción general desde la perspectiva de los usuarios en el contexto de investigación e innovación en ciencia y tecnología en el Departamento del Cauca. Para llevar a cabo la experimentación se definió realizar pruebas libres o “a ciegas”

para recolectar información sobre la experiencia percibida por los usuarios al usar el *Dashboard*, asegurando la objetividad y flexibilidad en la recopilación y análisis de resultados.

7.3.2. Selección del contexto

Los ambientes en donde se llevó a cabo la evaluación fueron las salas de cómputo de la Facultad de Ingeniería Electrónica y Telecomunicaciones, el salón de consejo de la Vicerrectoría de Investigaciones (*VRI*) en el Museo de Historia Natural, y en línea mediante videoconferencias. Para la evaluación se utilizó una encuesta descriptiva y exploratoria con el fin de capturar afirmaciones sobre las características de la herramienta e identificar problemas iniciales o bien nuevas funcionalidades que podrían ser agregadas en desarrollos futuros.

En los espacios compartidos se realizaron breves presentaciones del proyecto otorgando información sobre los objetivos de la investigación, la recolección de los datos, el funcionamiento de la herramienta, la escalabilidad del proyecto y un video de un ejemplo introductorio acompañado de las instrucciones como guías de usuario. Sumado a esto, se compartió la encuesta con los interesados y se hizo un acompañamiento constante para resolver las dudas y mantener diálogos de realimentación. De esta forma se definió la naturaleza del contexto.

7.3.3. Selección de sujetos

Para capturar una idea general del aporte que puede ofrecer el *Dashboard* para los interesados en el contexto del análisis bibliométrico de la actividad científica de los grupos de investigación en el Cauca, se eligieron personas vinculadas al ecosistema de ciencia, tecnología e innovación. Entre los sujetos que fueron seleccionados se encuentran perfiles tales como investigador, director de grupo de investigación, docente, administrador *VRI*, jefe de División de Gestión de la Investigación *VRI*, administrador División de Articulación con el Entorno (*DAE*), coordinador de Propiedad Intelectual *VRI*, codirector *ECoS-CTeI* y otros roles en *ECoS-CTeI*. Todos los sujetos cuentan con estudios profesionales o posgraduales, experiencias y conocimientos en el entorno de la investigación científica y la innovación.

Para esta investigación se encuestó a un total de 19 sujetos para la evaluación. Si bien es considerable señalar que es necesario un número elevado de personas para

ejecutar una evaluación correcta de la usabilidad de un *software,* un estudio demuestra que un gran porcentaje de las situaciones comprometedoras de una herramienta pueden ser reveladas en pruebas de usabilidad cuando la cifra de encuestados supera los 10, obteniendo excelentes resultados con valores de entre 15 y 20 individuos [81]. Adicionalmente, se hace énfasis sobre el valor que tiene el alto nivel de relación, experiencia y conocimientos de los sujetos seleccionados con respecto al contexto tratado.

7.3.4. Instrumentación

La adecuación del *Dashboard* y de los elementos utilizados para guiar a los participantes en la experimentación constaron de:

- ***Dashboard:*** Aplicación *web* que permite utilizar las funciones expuestas en la sección 6.4, dotada con la información extraída por los módulos extractores de los capítulos 4 y 5. Desplegada durante las jornadas de evaluación presenciales en las salas 334 y 331 de la *FIET,* y en una *IP* pública ofrecida temporalmente por Amazon.
- **Guías de usuario:** Exposición de la información necesaria para que los participantes puedan entender el contexto y funcionalidad de la herramienta, además del acompañamiento y dialogo constante durante las pruebas. También se comparte un video de 4 minutos con un ejemplo introductorio disponible en: https://youtu.be/tdp8FPcWMco
- **Encuesta:** Captura organizadamente información de la experiencia percibida por el usuario al usar el *Dashboard,* como también su opinión sobre la usabilidad de este. El cuestionario utilizado se compone de 18 preguntas con el motivo de evaluar las características y requisitos del *Dashboard*. Este se construyó con base en los procesos de evaluación realizados en propuestas de investigación similares a la del presente trabajo de grado [45], [49]. Para el formulario se utilizó la herramienta gratuita en línea de *Google Forms*. Con esto se estudió la reacción de los usuarios, la disposición de las secciones, terminología, gráficos e información, el aporte y utilidad percibido, el conocimiento previo requerido para el uso de tableros de análisis, la fluidez, la funcionalidad, las características redundantes o de interés a futuro, y los aspectos destacados y a mejorar. El cuestionario utilizado y las respuestas de la encuesta se pueden observar en el Anexo F.

7.3.5. Ejecución de pruebas

La ejecución inició con el contacto de sujetos y el acuerdo de reuniones presenciales y virtuales. Esto con ayuda del codirector y asesor del trabajo de grado. Durante este paso de la experimentación se vigiló constantemente a la herramienta para asegurar su correcto funcionamiento y se asistió a los diferentes espacios con los instrumentos necesarios para las presentaciones. También se mantuvieron diálogos de realimentación con los sujetos. Posteriormente, se hizo un seguimiento de las respuestas en la encuesta y del comportamiento del *Dashboard* en respuesta a las situaciones de estrés en el uso simultaneo. Finalmente se obtuvieron todos los cuestionarios diligenciados por los 19 sujetos que participaron de la experimentación.

7.4. ANÁLISIS DE RESULTADOS

La encuesta generó resultados variados para los 19 encuestados. Inicialmente se tomaron los nombres de los participantes, su área de trabajo y su rol principal en las 3 primeras preguntas del cuestionario. Se tuvo que todos ellos hacían parte de la academia y ninguno pertenecía al sector social o industrial. Con respecto a los roles principales se registraron 3 (15.8%) administradores, 8 (42.1%) investigadores y 8 (42.1%) profesionales equivalentes a las proporciones de la Figura 7.4. No se registraron estudiantes o roles alternativos a los mencionados. Todos los participantes se encontraban familiarizados en la materia del contexto, en mayor o menor medida, como se mencionó en la sección 7.3.3.

Figura 7.4. Resultados del cuestionario sobre los roles de los participantes. Fuente propia.

La evaluación de la experiencia general percibida por los participantes se resolvió mediante las respuestas de las preguntas 4 y 5 presentadas en la Figura 7.5. La apariencia general del *Dashboard* para más de la mitad de los sujetos fue “totalmente agradable”, para casi una tercera parte ”moderadamente agradable” y para la minoría restante no fue “ni agradable ni desagradable”, por lo que en mayor medida la reacción fue positiva y en menor medida neutral en cuanto a la estética y presentación. Con respecto al manejo del *Dashboard,* 18 (94.7%) de los

encuestados pensaron que era al menos “moderadamente fácil” y uno de ellos consideró que no era “ni fácil ni complicado”, destacando una gran percepción de facilidad en el uso de la herramienta. De estas respuestas se considera que la reacción es mayoritariamente positiva, sin embargo, las opiniones neutrales permiten evidenciar que algunos aspectos podrían ser mejorados. La síntesis de algunos comentarios particulares durante los diálogos de las pruebas se comparte al final del análisis de resultados en la Tabla 7.1 en adición a las últimas preguntas que fueron de respuesta libre.

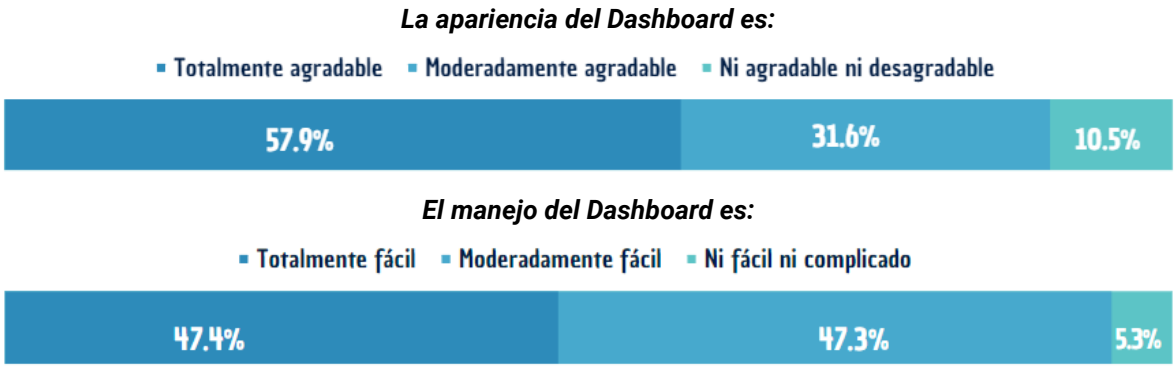


Figura 7.5. Resultados del cuestionario sobre la experiencia general percibida por los participantes durante el uso del *Dashboard*. Fuente propia

La organización e interpretación de la información gráfica y textual en las ventanas del *Dashboard* se evaluaron en las preguntas 6, 7 y 8 del cuestionario como se observa en la Figura 7.6. La organización de la información en las ventanas se consideró “totalmente clara” por 12 (63.2%) de los participantes, “moderadamente clara” por 5 (26.3%) de ellos y “moderadamente confusa” por 2 (10.5%) de ellos, notando una claridad mayoritaria para interpretar la disposición de la información presentada y, a su vez, pequeñas dificultades por una parte mínima de los participantes. Con este resultado se da lugar a mejoras en la disposición de la información para futuros desarrollos, o bien, a la consideración de capacitaciones enfocadas en facilitar la interpretación de los reportes generados. En cuanto a la información gráfica, 18 (94.7%) de los encuestados consideraron que esta era al menos “moderadamente fácil de interpretar” y 1 (5.3%) de ellos mantuvo una postura neutral considerando que no era “ni fácil ni difícil de interpretar”. Por otro lado, la información textual se consideró al menos “moderadamente fácil de leer” por 17 (89.4%) de los participantes, “ni fácil ni difícil de leer” por 1 (5.3%) y “moderadamente difícil de leer” por otro (5.3%). Estos resultados sugieren que las gráficas fueron casi en su totalidad fáciles de interpretar. Sin embargo, también

podrían darse mejoras ligeras sobre estas y sobre la información textual en desarrollos futuros facilitando su lectura para el usuario final al manejar un equilibrio entre tamaños y colores preferidos. Esto sin perder la satisfacción de la mayoría que lo encontró cómodo.

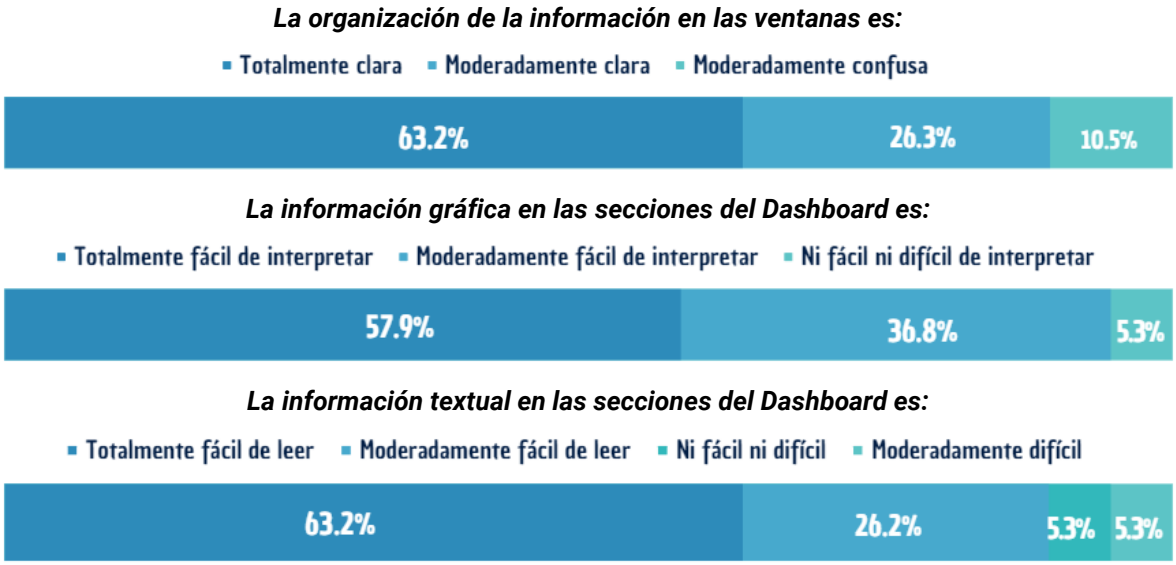


Figura 7.6. Resultados del cuestionario sobre la organización e interpretación de la información presentada. Fuente propia.

La funcionalidad y relevancia de las características principales del *Dashboard* se evalúan en las preguntas 9, 10 y 11, expuestas en la Figura 7.7. Las funciones de la herramienta fueron consideradas como “totalmente coherentes” por 12 (63.2%) de los participantes, “moderadamente coherentes” por 6 (31.5%) de ellos y “ni coherentes ni incoherentes” por 1 (5.3%). El proceso de filtrado de la información y la generación de los reportes presentó resultados similares siendo considerado “totalmente consistente” para 11 (57.9%) participantes, “moderadamente consistente” para 7 (36.8%) y “ni consistente ni inconsistente” para 1 (5.3%). Por otro lado, todos los encuestados opinaron que los indicadores bibliométricos utilizados eran al menos “moderadamente relevantes”. Con esto se tiene que los resultados de este apartado fueron casi totalmente positivos y mínimamente neutrales, destacando la satisfacción por parte de los participantes con respecto a las capacidades y valor del *Dashboard*. Aun así, existen aspectos que pueden perfeccionarse.

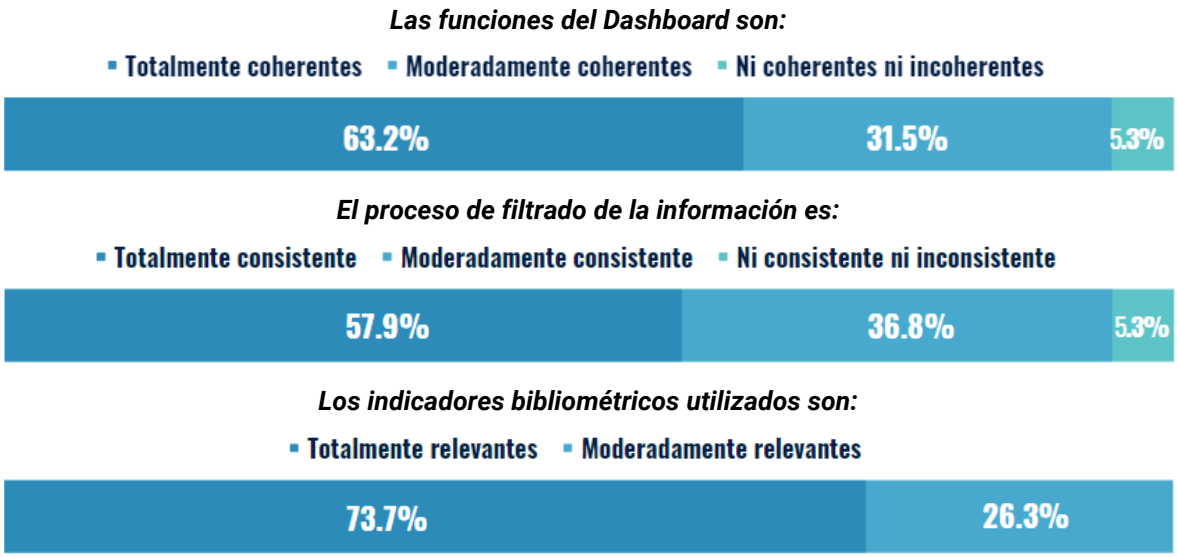


Figura 7.7. Resultados del cuestionario sobre la funcionalidad y relevancia del *Dashboard*. Fuente propia.

La terminología del *Dashboard* se evaluó en la pregunta 12 del cuestionario, expuesta en la Figura 7.8. Todos los encuestados opinaron que el lenguaje utilizado en la herramienta era al menos “moderadamente adecuado”, siendo mayoría los que lo consideraron “totalmente adecuado”. De esta forma la terminología usada solo generó respuestas positivas.

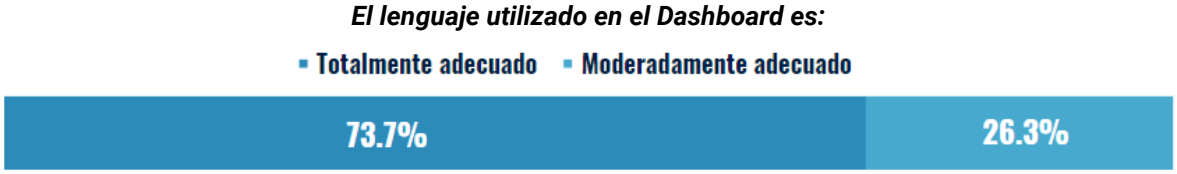


Figura 7.8. Resultados del cuestionario sobre el lenguaje utilizado en el *Dashboard.* Fuente propia.

Mediante las preguntas 13 y 14, se evaluó el nivel de conocimiento previo requerido en el uso de tableros de análisis y que tan intuitivo fue el uso de la herramienta presentada. Los resultados muestran variaciones considerables en la Figura 7.9. Con respecto a si los usuarios podrían usar el *Dashboard* sin instrucciones, casi la mitad refirieron que estaban “totalmente de acuerdo”, el restante se dividió entre opiniones de “moderadamente de acuerdo”, “moderadamente en desacuerdo” y “totalmente en desacuerdo”, siendo sólo 4 (21%) usuarios los que expresaron un grado de dificultad ante la situación hipotética de no tener disponibles las instrucciones. Sobre la necesidad de tener conocimiento o experiencia previa en el uso de tableros de análisis, casi la mitad de los usuarios expresó que era

“moderadamente necesario” y la otra mitad se dividió entre “totalmente necesario”, “moderadamente innecesario” y “ni necesario ni innecesario”, siendo sólo 4 usuarios los que pensaron, a mayor o menor grado, que era innecesario contar con experiencia previa. Dado que 15 (79%) de los usuarios expresaron que podrían usar la herramienta sin instrucciones y 14 (73.7%) consideró necesario tener algún tipo de conocimiento previo en el uso de tableros de análisis, se tiene que el uso de la herramienta fue bastante intuitivo, pero se hace necesario llevar a cabo una capacitación, o bien construir una guía o tutorial más exhaustivo con el objetivo de compensar la necesidad percibida de haber usado antes tableros de análisis. Esto con tal de facilitar el manejo y entendimiento del *Dashboard*.

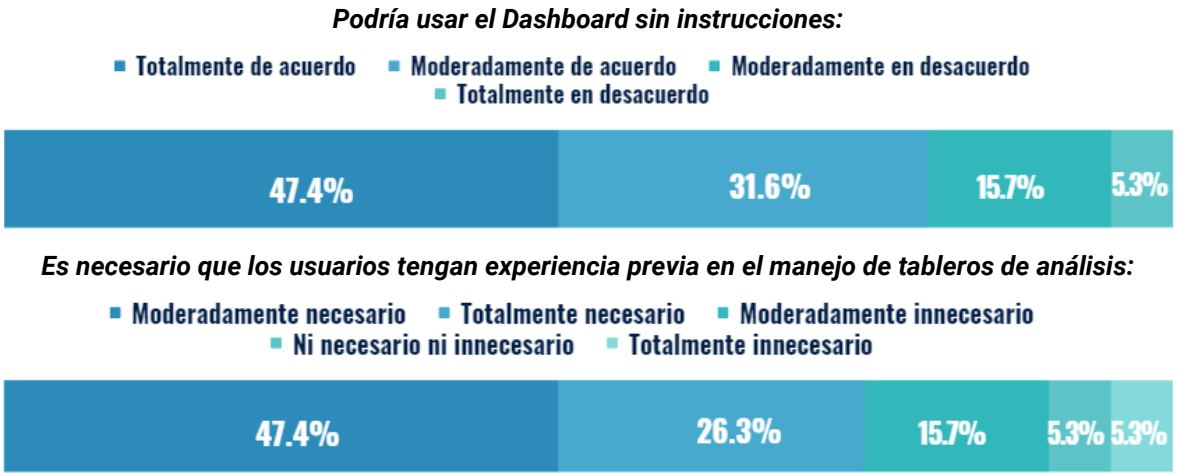


Figura 7.9. Resultados del cuestionario sobre el conocimiento previo requerido para el uso del *Dashboard*. Fuente propia.

La última pregunta de opción múltiple del cuestionario se refirió al aporte percibido por los participantes con respecto a la utilidad de la herramienta en el contexto del análisis bibliométrico de los grupos de investigación en el Cauca. De esta forma se obtuvo una opinión general sobre las capacidades de la herramienta propuesta de frente a los objetivos de la presente investigación. La Figura 7.10 presenta los resultados registrados destacando una gran mayoría de usuarios (14 o 73.7%) que consideraron que el análisis y visualización bibliométrica de los grupos de investigación ”siempre” puede ayudar a los actores del *SRCTI*. De entre los usuarios restantes, 3 (15.8%) consideraron que esto puede aplicarse “muchas veces” y 2 (10.5%) opinaron que “algunas veces”. Con esto se logró identificar la importancia que puede adquirir el *Dashboard* al ser aplicado por diferentes personas del entorno investigativo en la región.

El análisis y visualización bibliométrica de los grupos de investigación puede ayudar a los actores del Sistema Regional de Ciencia, Tecnología e Innovación del Cauca:

Figura 7.10. Resultados del cuestionario sobre la posible ayuda generada por el *Dashbaord* a los actores del SRCTI. Fuente propia.

En última instancia, se presenta la Tabla 7.1 la cual refleja las respuestas a las preguntas libres del cuestionario enfocadas en capturar los aspectos más y menos gustados, así como características que podrían ser incluidas según el interés particular y perspectiva de los sujetos encuestados si se desarrollaran futuras versiones de la aplicación. Cabe mencionar que también se tuvieron en cuenta las ideas de los diálogos compartidos que no quedaron del todo plasmadas en la encuesta, pero que de alguna manera llamaron la atención de los usuarios. La síntesis presentada a continuación puede servir como base para la creación de una versión de producción o trabajo a futuro que extienda la presente investigación.

	Aspectos destacados	Aspectos para mejorar
Análisis y visualización bibliométrica	- Diversidad de fuentes - Análisis comparativo de los grupos - Análisis individual de los grupos - Información estratégica y relevante para toma de decisiones - Información exportable	- Detalles de la información visible en elementos particulares de las gráficas generadas - Diferenciación de elementos avalados y no avalados - Opción de variar los tipos de gráficos utilizados - Normalización de variaciones de nombres de autor más avanzada - Listado de los resultados obtenidos - Inclusión de más tipos de productos - Macro categorías
Exploración de datos	- Acceso eficiente a la información bibliográfica - Novedoso	- Combinar varios conjuntos de criterios de filtrado en una misma búsqueda - Potenciar funcionalidad de las tablas de información - Los resultados de la opción “Todos” no siempre conduce a resultados relevantes - Exportación de tablas de información
Extracción de datos	- Actualizar datos - Escalabilidad	- Poner a pie de página la última fecha de actualización de datos

Apariencia general de la herramienta	- Organización de los datos - Interfaz cómoda a la vista	- Estética de la tabla de información - Texto pequeño en los diagramas - Algunos términos no son tan fáciles de entender - Diseño *front-end* - Mensajes de ayuda sobre cada elemento presentado
Usabilidad de la herramienta	- Fluidez entre procesos y transiciones - Facilidad de uso de las funciones	- Asesoría, capacitación o guía previa al uso
Otros elementos	- Impacto para procesos de acreditación	- Predicción para la clasificación de grupos

Tabla 7.1. Síntesis de aspectos señalados por los encuestados sobre el *Dashboard*. Fuente propia.

8. DISCUSIÓN

En este capítulo se comparten una serie de hallazgos encontrados durante la construcción de los módulos y el desarrollo de la investigación. Así mismo, se resaltan las conclusiones de mayor importancia sobre la línea de los objetivos planteados, las contribuciones logradas y por último se proponen distintos trabajos a futuro que pueden potenciar los resultados obtenidos para fortalecer el Sistema Regional de Ciencia, Tecnología e Innovación.

8.1. HALLAZGOS

Durante las etapas de integración, emparejamiento y limpieza de datos entre las fuentes de *CVLAC, GrupLAC* y *Scopus,* descritas en las secciones 5.3.3, 6.1.3 y 6.4.1 de la monografía, se identificaron varios hallazgos a partir de diversas estadísticas y condiciones de los conjuntos de datos en general. Los reportes de estos valores pueden ser recolectados a través de la consola o terminal durante la ejecución de la etapa de preprocesamiento del sistema. Los valores están sujetos a cambios según se actualicen las bases de datos. Por tanto, las cifras presentadas a continuación están actualizadas a la fecha de 10 de Julio de 2023. Los datos atribuidos a Minciencias provienen de la herramienta de “La ciencia en cifras – Grupos de investigación reconocidos“ [10], [12].

- El total de investigadores identificados en el Cauca por Minciencias fue de 370. Esta investigación arroja que la cifra fue de 3062 para *GrupLAC* y 2201 para *Scopus*, evidenciando una abundante diferencia respecto a las cifras oficiales.
- Los datos emparejados en *Scopus* para el Cauca, permitieron identificar los grupos de investigación a los que pertenecen 181 investigadores entre un total de 2201 para la base de datos generada, de los cuales al menos 306 cuentan con un identificador de autor de *Scopus* registrado. Esto es evidencia de que una alta tasa de investigadores (86%) no registraron su identificador *Scopus* en sus hojas de vida de *CVLAC*, otra parte lo hizo erróneamente (6%) y el restante lo ingresó correctamente (8%).
- Los datos emparejados en *Scopus* para el Cauca, permitieron identificar los grupos de investigación a los que pertenecen 1646 productos entre un total de 3067 para la base de datos generada, evidenciando que poco menos de

la mitad de estos no están registrados en la plataforma de GrupLAC, se registraron erróneamente o presentan información incompleta para su emparejamiento.

- El total de grupos de investigación identificados en el Cauca por Minciencias fue de 117. Para la extracción realizada desde *GrupLAC* la cifra fue de 118 de los cuales 114 son visibles en *Scopus* a partir de algunos autores y productos. Esto es evidencia de que se puede rastrear la actividad científica de los grupos en repositorios de alto impacto de manera adecuada teniendo en cuenta que la calidad del análisis dependerá de la calidad de los datos registrados y los métodos con los que estos son capturados.
- Los datos extraídos de *GrupLAC* permitieron identificar que, tan sólo respecto a artículos de investigación, al menos 62 de entre 118 grupos de investigación presentan duplicados en su perfil. Incluso artículos verificados por Minciencias presentan esta condición, siendo al menos 56 los grupos con artículos a la vez verificados y duplicados en su propio apartado del perfil *GrupLAC*. Igualmente, se identificaron al menos 873 artículos duplicados dentro de los propios perfiles de los grupos, evidenciando una situación de sesgo en los procesos de categorización que usan métricas “verificadas” a modo de indicadores, pues son susceptibles a inflarse artificialmente. Se llegó a identificar grupos con hasta 7 veces el mismo artículo registrado y verificado en su perfil, con título y *DOI* exacto.
- De una lista inicial de 131 afiliaciones, hecha a partir de las *APIs* de *Scopus,* se obtuvieron 64 afiliaciones tras la limpieza de datos. Esto es evidencia de que existen perfiles de afiliación en *Scopus* falsos, residuales o que han sido mal gestionados por algunas instituciones, dividiendo su información bibliográfica y resultando contraproducente para la medida de su impacto y productividad. Una situación similar se presenta con perfiles múltiples creados por autores.
- Existen variables categóricas en *GrupLAC,* referentes a la actividad científica, que brindan poco o nulo conocimiento de la materia en cuestión. Un ejemplo de esto son las líneas de investigación, de las cuales se registran 531 líneas diferentes en el Cauca, siendo sólo 10 de ellas compartidas por un máximo de 2 grupos de investigación. Esto es evidencia de que la gran mayoría de líneas de investigación no representan categorías como tal, si no información libre que es definida por los propios grupos.

Los hallazgos sugeridos permiten señalar varias situaciones importantes que permean desde la naturaleza de la captación de información bibliográfica, hasta la calidad de los análisis que pueden resultar de las herramientas enfocadas. Además, las condiciones estudiadas implican la acción tanto de los individuos como de las entidades que pertenecen al entorno. Se recomienda incentivar buenas prácticas sobre el manejo de la información bibliográfica veraz, precisa y completa, como también fortalecer los recursos dispuestos para capturar, analizar y concluir aspectos estratégicos en la toma de decisiones que deriva de esta información.

8.2. CONCLUSIONES

En la presente investigación se implementó un *Dashboard* para el análisis y visualización bibliométrica dentro del ámbito de los grupos de investigación científica en el Cauca, basándose en la extracción y persistencia de datos de plataformas de *ScienTI* y la base de datos de *Scopus.* La herramienta se desarrolló con el fin de contribuir a la conducción de una mejor toma de decisiones por parte de los actores del Sistema Regional de Ciencia, Tecnología e Innovación. En virtud de esa idea, se modeló, construyó y evaluó un conjunto de módulos integrados que buscan ser una herramienta de apoyo para todos los interesados en la extracción de conocimiento sobre los grupos de investigación en la región. Del mismo modo, se plasmaron las contribuciones alcanzadas con el desarrollo del trabajo de grado y el análisis de los resultados y hallazgos obtenidos. A continuación, se presentan las conclusiones generales de la investigación:

- En el estado actual del conocimiento se observaron propuestas para el análisis bibliométrico de la información de las bases de datos bibliográficas de mayor reconocimiento en el mundo, así como de otras fuentes menos reconocidas, pero mejor relacionadas con el contexto nacional y regional. No obstante, no se identificó una herramienta completa en aspectos de recolección, actualización, granularidad y exploración de la información referente a la población objetivo. Consecuentemente, esta investigación cobra relevancia en el ámbito de la bibliometría y cienciometría caucana al proponer una herramienta que comprende etapas del flujo de datos desde su captura hasta su análisis, notando igualmente capacidades novedosas para su escalabilidad.
- La calidad del análisis y visualización bibliométrica de los grupos de investigación depende directamente de la calidad y abundancia de los datos

que pueden recolectarse. La abstención del registro de los datos, su ingreso de manera descuidada o incompleta y la falta de rigurosidad de los recursos principales u oficiales para su gestión en el contexto colombiano, limitan el estudio del impacto y desarrollo del ecosistema investigativo en sí como resultado de la articulación de sus partes. Si bien se pueden aplicar ciertas técnicas para compensar las situaciones mencionadas, el sesgo presente en la toma de decisiones y planeación estratégica que deriven de este tipo de análisis, será inversamente proporcional al grado de fiabilidad que otorguen los modos de gestión de la información bibliográfica por parte de individuos y entidades.

- El acceso a la información de la investigación científica actualizada y completa en Colombia, considerando las plataformas oficiales, es tedioso y complicado. En ese sentido, los conjuntos de datos generados en este trabajo de investigación, junto con los medios usados para su creación, incentivan el suministro regulado y eficiente de datos abiertos y organizados para personas con o sin conocimiento técnico. Estos medios de suministro además son extensibles a otros territorios y a la construcción de más soluciones analíticas particulares a partir de las múltiples dimensiones de información accesible para lograr más y mejores investigaciones o desarrollos en áreas relacionadas.
- El *Dashboard* construido permite utilizar las fuentes *CVLAC*, *GrupLAC* y *Scopus* para extraer conocimiento de la actividad, productividad, colaboración, participación e impacto de los grupos de investigación en el Cauca por parte de los actores del *SRCTI*. En ese sentido, se propone como una herramienta interesante para apoyarlos en el mejoramiento de la toma de decisiones particulares según el rol de cada actor, desde estrategias para la dirección de un grupo de investigación hasta la administración de recursos o reconocimientos, entre otros. Adicionalmente, esta puede ser nutrida con más datos, filtros, gráficas o elementos complementarios.
- La evaluación del prototipo final propuesto fue un éxito, así como los métodos de verificación y validación orientados a asegurar el correcto funcionamiento de los módulos extractores. Por un lado, se contó con ingenieros expertos de la *FIET* para la inspección individual de cada requisito de los módulos extractores en dos sesiones de validación separadas y subsecuentes a procesos de verificación para cada una de las herramientas con resultados positivos. Por otro lado, se contó con un grupo de 19 expertos, todos altamente relacionados al entorno investigativo del Cauca y varios de ellos

con cargos importantes en la *VRI*, *DAE* y *ECoS-CTeI.* Este grupo se encargó de la evaluación del *Dashboard* mediante una encuesta y diversos diálogos mantenidos en espacios compartidos. Los resultados demostraron ser positivos en términos de usabilidad y satisfacción por parte de los encuestados, como también permitieron señalar aspectos destacados y a mejorar con respecto a su funcionalidad.

8.3. TRABAJOS A FUTURO

Teniendo en cuenta las oportunidades de investigación que se abren con el desarrollo de este trabajo de grado, se proponen los siguientes trabajos a futuro:

- **Extender el análisis y visualización bibliométrica de los grupos de investigación a nivel nacional:** este proyecto ha sido diseñado con una arquitectura escalable que permite expandir el alcance desde el Departamento del Cauca a otros departamentos del país. Dado a su estructura flexible y la optimización del rendimiento implementada, se propone adaptar más listas iterables de los grupos de investigación en el sistema a partir de los buscadores de Minciencias con el objetivo de estudiar nuevos enfoques y alimentar las bases de datos. Proyectos de mayor alcance al planteado en esta investigación requieren de mayores recursos de cómputo que los utilizados para el procesamiento y persistencia de los datos. Así mismo, se sugiere considerar otros aspectos importantes como permisos, repercusión y contingencias técnicas. La exploración, análisis y visualización bibliométrica de los datos que ofrece el *Dashboard* es adaptable a nuevos registros y conjuntos de datos. No obstante, altos volúmenes de información con respecto a los niveles manejados en este proyecto aún no han sido probados. Finalmente, cabe mencionar que también es posible construir y agregar nuevos algoritmos de extracción de datos que correspondan a tablas adicionales de información.
- **Desarrollar nuevas versiones del prototipo encaminadas a enfoques particulares de analítica de datos:** si bien el sistema presentó el funcionamiento esperado, los procesos de evaluación pusieron en evidencia los detalles y variedad de las necesidades percibidas por los diferentes roles que tienen los actores del *SRCTI*. El análisis de resultados de esta investigación permitió conocer intereses particulares a futuro dirigidos a diferentes líneas de usabilidad enfocadas en funcionalidades adicionales

sobre la exploración de datos, la personalización del análisis visual, las opciones de filtrado, la inclusión de más tipos de productos, entre otros. Con base en la literatura científica, también se propone fortalecer el mapeo científico entre los grupos de investigación construyendo redes de colaboración más exhaustivas. Esto puede aportar un mayor valor al análisis en especial si tienden a crecer los conjuntos de datos utilizados. Por último, es recomendable la inclusión de un mayor número de expertos e interesados que lleven a cabo nuevas etapas de evaluación de la herramienta para estudiar los aportes de mayor beneficio.

- **Desarrollar un sistema de análisis bibliométrico predictivo para los grupos de investigación:** un enfoque interesante ante las nuevas tecnologías emergentes y la evolución constante y vertiginosa de los modelos predictivos, se propone mediante el desarrollo de un sistema de análisis bibliométrico capaz de pronosticar el comportamiento o desempeño de los grupos de investigación basado en las principales variables de estudio que determinan las métricas o indicadores más utilizados en la actualidad según la literatura científica. Mediante técnicas de aprendizaje profundo, aprendizaje automático, modelamiento matemático y modelamiento estadístico es posible construir, sobre las bases del presente trabajo de grado, nuevos proyectos e investigaciones dirigidas a prever el impacto de las publicaciones, identificar tendencias anticipadamente, optimizar la gestión de recursos, implementar analítica prescriptiva y, en general, aportar al mejoramiento de la toma de decisiones estratégica en el ámbito científico de los grupos en la región y el país.
- **Construir un conjunto de datos de la actividad científica de los grupos de investigación apoyado en "*Linked Open Data*":** las nuevas tendencias de acceso a la información, basadas en la *web* semántica, permiten el enlazamiento de datos abiertos y estructurados para ser aprovechados en el ámbito global de la *web* mediante la contextualización sofisticada de estos a partir de sus vocablos y conceptos. Ahora bien, la construcción de modelos conceptuales basados en datos abiertos y enlazados de la actividad científica de los grupos de investigación en el Cauca y en Colombia, se propone como un trabajo a futuro. Esto debido a su potencial para contribuir a nuevos desarrollos de aplicaciones o servicios que permitan la extracción de conocimiento de valor académico de manera libre y gradual. Por ejemplo, la construcción y consumo compartido de mapas de investigación científica con las ventajas que ofrece la *web* semántica.

9. BIBLIOGRAFÍA

[1] M. Otero y R. Alejandro, «La investigación como elemento fundamental para el desarrollo de Latinoamérica. Tendencias y perspectivas», *Rev. CIECAS-IPN*, n.º 39, pp. 35-44, 2016.

[2] ECoS-CTeI, «ECos-CTeI». https://vri.unicauca.edu.co/ecos-ctei/ (accedido 14 de marzo de 2023).

[3] Ministerio de Ciencia Tecnología e Innovación, «Acerca de la Red SCienTI : ScienTI». http://www.scienti.net/php/level.php?lang=es&component=19&item=1 (accedido 24 de marzo de 2022).

[4] Ministerio de Ciencia Tecnología e Innovación, «Plataforma SCIENTI - Colombia», *Minciencias*. https://minciencias.gov.co/scienti (accedido 30 de julio de 2023).

[5] Clarivate, «Web of Science platform», *Clarivate*. https://clarivate.com/products/scientific-and-academic-research/research-discovery-and-workflow-solutions/webofscience-platform/ (accedido 30 de julio de 2023).

[6] Elsevier, «About Scopus - Abstract and citation database | Elsevier». https://www.elsevier.com/solutions/scopus (accedido 24 de marzo de 2022).

[7] Gobernación del Cauca, Colciencias, Banco Interamericano de Desarrollo, Universidad del Cauca, y CODECTI, *Plan Estratégico Departamental de Ciencia, Tecnología e Innovación del Cauca*. Universidad del Cauca, 2013. [En línea]. Disponible en: http://hdl.handle.net/20.500.12324/34596

[8] D. Hicks, P. Wouters, L. Waltman, S. de Rijcke, y I. Rafols, «Bibliometrics: The Leiden Manifesto for research metrics», *Nature*, vol. 520, n.º 7548, pp. 429-431, abr. 2015, doi: 10.1038/520429a.

[9] Región Administrativa y de Planificación del Pacífico, «Sinergia universidad-empresa-estado-sociedad civil como estrategia de innovación para el desarrollo regional». Rap Pacifico, 2020. [En línea]. Disponible en: https://rap-pacifico.gov.co/wp-content/uploads/2020/07/WEBINAR-II.pdf

[10] Ministerio de Ciencia Tecnología e Innovación, «La Ciencia en Cifras», *Minciencias*. https://www.minciencias.gov.co/la-ciencia-en-cifras (accedido 30 de julio de 2023).

[11] Ministerio de Ciencia Tecnología e Innovación, «Estadísticas Generales Grupos e Investigadores». https://minciencias.gov.co/la-ciencia-en-cifras/estadisticas-generales (accedido 13 de febrero de 2023).

[12] Ministerio de Ciencia Tecnología e Innovación, «Grupos de Investigación reconocidos por Minciencias», *Minciencias*. https://minciencias.gov.co/la-ciencia-en-cifras/grupos (accedido 16 de febrero de 2023).
[13] Ministerio de Ciencia Tecnología e Innovación., «CvLAC». https://scienti.minciencias.gov.co/cvlac (accedido 30 de julio de 2023).
[14] Ministerio de Ciencia Tecnología e Innovación, «Manual de usuario CvLAC». 6 de julio de 2021. [En línea]. Disponible en: https://minciencias.gov.co/sites/default/files/ckeditor_files/D103M06%20Manual%20de%20Usuario%20CVLAC%20V01do.pdf
[15] Ministerio de Ciencia Tecnología e Innovación, «GrupLAC». https://scienti.minciencias.gov.co/gruplac/ (accedido 30 de julio de 2023).
[16] Ministerio de Ciencia Tecnología e Innovación, «Manual de usuario GrupLAC». 6 de julio de 2021. [En línea]. Disponible en: https://minciencias.gov.co/sites/default/files/ckeditor_files/D103M03%20Manual%20GrupLAC%20V01%20do.pdf
[17] O. Calvo Giraldo, «La Gestión del Conocimiento en las Organizaciones y las Regiones: Una Revisión de la Literatura», *Tendencias*, vol. 19, n.º 1, pp. 140-163, jul. 2018, doi: 10.22267/rtend.181901.91.
[18] L. I. Tello Clavijo, «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, Bogotá, Colombia, 2019. [En línea]. Disponible en: https://repositorio.unal.edu.co/bitstream/handle/unal/75992/1110537086.2019.pdf
[19] D. M. Mosquera Echeverry, G. De la Torre Solarte, G. Bastidas Gustín, O. Calvo Giraldo, y S. M. Sandoval Ruíz, «Construyendo redes para potenciar la innovaciòn en el Cauca». 2019. [En línea]. Disponible en: http://www.unicauca.edu.co/innovaccioncauca/node/4141
[20] ECoS-CTeI, «Boletin ECoS-CTeI Enero 2023». 2023. Accedido: 14 de marzo de 2023. [En línea]. Disponible en: https://drive.google.com/file/d/1P9YVempXNl2_w12iL56efx473d8Pn-o1/view
[21] Ministerio de Ciencia, Tecnología e Innovación, «Convocatoria nacional para el reconocimiento y medición de grupos de investigación, desarrollo tecnológico o de innovación y para el reconocimiento de investigadores del sistema nacional de ciencia, tecnología e innovación». 2021. [En línea]. Disponible en: https://minciencias.gov.co/sites/default/files/upload/convocatoria/anexo_1_-_documento_conceptual_2021.pdf

[22] Colciencias, «Ministerio de Ciencia Tecnología e Innovación». [En línea]. Disponible en: https://minciencias.gov.co/sites/default/files/upload/glosario-colciencias.pdf
[23] Ministério de Ciência e Tecnologia e Inovação, «Plataforma Lattes». https://lattes.cnpq.br/ (accedido 14 de marzo de 2023).
[24] CNPq, «Extração de dados - Portal Memória». https://memoria.cnpq.br/web/portal-lattes/extracoes-de-dados (accedido 28 de julio de 2023).
[25] Elsevier, «Scopus fact sheet 2022». 2022. [En línea]. Disponible en: https://www.elsevier.com/__data/assets/pdf_file/0017/114533/Scopus-fact-sheet-2022_WEB.pdf
[26] K. Wan, «What are Scopus APIs and how are these used?» [En línea]. Disponible en: https://www.elsevier.com/__data/assets/pdf_file/0007/917179/Scopus-User-Community-Germany-API-final.pdf
[27] Elsevier, «Elsevier Developer Portal». https://dev.elsevier.com/ (accedido 14 de marzo de 2023).
[28] Elsevier, «API Interface Specification». https://dev.elsevier.com/api_docs.html (accedido 16 de febrero de 2023).
[29] Elsevier, «Abstract Retrieval API», *Abstract Retrieval API*. https://dev.elsevier.com/documentation/AbstractRetrievalAPI.wadl (accedido 16 de febrero de 2023).
[30] Elsevier, «Author Search API», *Author Search API*. https://dev.elsevier.com/documentation/AuthorSearchAPI.wadl (accedido 16 de febrero de 2023).
[31] Elsevier, «Author Retrieval API», *Author Retrieval API*. https://dev.elsevier.com/documentation/AuthorRetrievalAPI.wadl (accedido 16 de febrero de 2023).
[32] Elsevier, «Default API Key Settings». https://dev.elsevier.com/api_key_settings.html (accedido 14 de marzo de 2023).
[33] Elsevier, «Scopus Search API». https://dev.elsevier.com/documentation/ScopusSearchAPI.wadl (accedido 14 de marzo de 2023).
[34] Elsevier, «Affiliation Search API». https://dev.elsevier.com/documentation/AffiliationSearchAPI.wadl (accedido 14 de marzo de 2023).

[35] R. N. Broadus, «Toward a definition of “bibliometrics”», *Scientometrics*, vol. 12, n.º 5, pp. 373-379, nov. 1987, doi: 10.1007/BF02016680.
[36] V. P. Diodato y P. Gellatly, *Dictionary of Bibliometrics*. New York: Routledge, 1994. doi: 10.4324/9780203714133.
[37] A. Pritchard, «Statistical Bibliography or Bibliometrics?», *J. Doc.*, vol. 25, pp. 348-349, ene. 1969.
[38] J. Mingers y L. Leydesdorff, «A review of theory and practice in scientometrics», *Eur. J. Oper. Res.*, vol. 246, n.º 1, pp. 1-19, oct. 2015, doi: 10.1016/j.ejor.2015.04.002.
[39] R. Mitchell, *Web Scraping with Python: Collecting Data from the Modern Web*. Sebastopol, CA: O’Reilly, 2015.
[40] K. Pauwels *et al.*, «Dashboards as a Service: Why, What, How, and What Research Is Needed?», *J. Serv. Res.*, vol. 12, n.º 2, pp. 175-189, nov. 2009, doi: 10.1177/1094670509344213.
[41] Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU), «Coronavirus COVID-19 Dashboard», 2019. https://www.arcgis.com/apps/dashboards/bda7594740fd40299423467b48e9ecf6 (accedido 14 de marzo de 2023).
[42] A. Silberschatz, H. F. Korth, y S. Sudarshan, *Database system concepts*, 6th ed. New York: McGraw-Hill, 2011.
[43] M. Lorenz, G. Hesse, y J.-P. Rudolph, «Object-relational Mapping Revised - A Guideline Review and Consolidation»:, en *Proceedings of the 11th International Joint Conference on Software Technologies*, Lisbon, Portugal: SCITEPRESS - Science and Technology Publications, 2016, pp. 157-168. doi: 10.5220/0005974201570168.
[44] B. Kitchenham, «Procedures for Performing Systematic Reviews», *Keele UK Keele Univ*, vol. 33, ago. 2004.
[45] A. Dattolo y M. Corbatto, «Assisting researchers in bibliographic tasks: A new usable, real-time tool for analyzing bibliographies», *J. Assoc. Inf. Sci. Technol.*, vol. 73, n.º 6, pp. 757-776, 2022, doi: 10.1002/asi.24578.
[46] V. H. Patil, S. A. Bhavsar, y A. H. Patil, «An efficient author information retrieval tool for bibliographic record analysis», *J. Intell. Fuzzy Syst.*, vol. 39, n.º 1, pp. 341-353, 2020, doi: 10.3233/JIFS-191289.
[47] J. Ruiz-Rosero, G. Ramirez-Gonzalez, y J. Viveros-Delgado, «Software survey: ScientoPy, a scientometric tool for topics trend analysis in scientific publications», *Scientometrics*, vol. 121, n.º 2, pp. 1165-1188, nov. 2019, doi: 10.1007/s11192-019-03213-w.

[48] M. Aria y C. Cuccurullo, «bibliometrix: An R-tool for comprehensive science mapping analysis», *J. Informetr.*, vol. 11, n.º 4, pp. 959-975, nov. 2017, doi: 10.1016/j.joi.2017.08.007.

[49] M. Cobo, A. G. López-Herrera, E. Herrera-Viedma, y F. Herrera, «SciMAT: A new science mapping analysis software tool», *J. Am. Soc. Inf. Sci. Technol.*, vol. 63, pp. 1609-1630, ago. 2012, doi: 10.1002/asi.22688.

[50] P. S. Escarassatti y H. H. Bíscaro, «Visual representation of bibliographic production data from Lattes Platform», In Review, preprint, may 2022. doi: 10.21203/rs.3.rs-1665862/v1.

[51] J. P. Mena-Chalco y R. M. C. Junior, «scriptLattes: an open-source knowledge extraction system from the Lattes platform», *J. Braz. Comput. Soc.*, vol. 15, n.º 4, pp. 31-39, dic. 2009, doi: 10.1007/BF03194511.

[52] A. D. Alves, H. H. Yanasse, y N. Y. Soma, «LattesMiner: a multilingual DSL for information extraction from lattes platform», en *Proceedings of the compilation of the co-located workshops on DSM'11, TMC'11, AGERE! 2011, AOOPES'11, NEAT'11, & VMIL'11*, en SPLASH '11 Workshops. New York, NY, USA: Association for Computing Machinery, oct. 2011, pp. 85-92. doi: 10.1145/2095050.2095065.

[53] D. F. Galeano Durán y L. C. Prada Pérez, «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, Bogotá, Colombia, 2019. Accedido: 16 de mayo de 2022. [En línea]. Disponible en: http://repository.udistrital.edu.co/handle/11349/22537

[54] Project Management Institute, Ed., *The standard for project management and a guide to the project management body of knowledge (PMBOK guide)*, Seventh edition. Newtown Square, Pennsylvania: Project Management Institute, Inc, 2021.

[55] Ministerio de Ciencia Tecnología e Innovación, Juan Carlos Corrales Muñoz, «CvLAC - plataforma SCienTI - Colombia», *Hoja de Vida*. https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0000013021 (accedido 16 de febrero de 2023).

[56] Ministerio de Ciencia Tecnología e Innovación, «GrupLAC - plataforma SCienTI - Colombia», *Grupo de Ingeniería Telemática*. https://scienti.minciencias.gov.co/gruplac/jsp/visualiza/visualizagr.jsp?nro=00000000008160 (accedido 16 de febrero de 2023).

[57] Ministerio de Ciencia Tecnología e Innovación, «Busqueda de Grupos por Departamento: Cauca», *Ciencia y Tecnología para Todos*.

https://scienti.minciencias.gov.co/ciencia-war/busquedaGrupoXDepartamentoGrupo.do?sglPais=COL&sgDepartamento=CA (accedido 16 de febrero de 2023).
[58] Pallets, «Welcome to Flask — Flask Documentation (2.2.x)». https://flask.palletsprojects.com/en/2.2.x/ (accedido 5 de junio de 2023).
[59] The PostgreSQL Global Development Group, «PostgreSQL», *PostgreSQL*, 5 de junio de 2023. https://www.postgresql.org/ (accedido 5 de junio de 2023).
[60] SQLAlchemy, «SQLAlchemy». https://www.sqlalchemy.org (accedido 5 de junio de 2023).
[61] Leonard Richardson Revision, «Beautiful Soup Documentation». https://beautiful-soup-4.readthedocs.io/en/latest/ (accedido 5 de junio de 2023).
[62] NumFOCUS, «Pandas Documentation». https://pandas.pydata.org/docs/ (accedido 5 de junio de 2023).
[63] Python Software Foundation, «re — Regular expression operations», *Python documentation*. https://docs.python.org/3/library/re.html (accedido 5 de junio de 2023).
[64] Elsevier, «Scopus Affiliation Search Views». https://dev.elsevier.com/sc_affil_search_views.html (accedido 14 de marzo de 2023).
[65] Elsevier, «Scopus Author Search Views». https://dev.elsevier.com/sc_author_search_views.html (accedido 14 de marzo de 2023).
[66] Elsevier, «Scopus Search Views». https://dev.elsevier.com/sc_search_views.html (accedido 14 de marzo de 2023).
[67] Elsevier, «Scopus Abstract Retrieval Views». https://dev.elsevier.com/sc_abstract_retrieval_views.html (accedido 14 de marzo de 2023).
[68] Elsevier, «Scopus Author Retrieval Views». https://dev.elsevier.com/sc_author_retrieval_views.html (accedido 14 de marzo de 2023).
[69] Elsevier, «Elsevier Scopus APIs». https://dev.elsevier.com/sc_apis.html (accedido 14 de marzo de 2023).
[70] Elsevier, «About Elsevier». https://www.elsevier.com/about (accedido 14 de marzo de 2023).
[71] G. Prathap, «Hirsch-type indices for ranking institutions' scientific research output [4]», *Curr. Sci.*, vol. 91, oct. 2006.

[72] L. Colledge, *Snowball Metrics Recipe Book*, 3.ª ed. 2016. [En línea]. Disponible en: https://www.elsevier.com/research-intelligence/resource-library/snowball-metrics-recipe-book

[73] Elsevier, «Metrics | Snowball Metrics». https://snowballmetrics.com/metrics/ (accedido 16 de febrero de 2023).

[74] J. E. Hirsch, «An index to quantify an individual's scientific research output», *Proc. Natl. Acad. Sci.*, vol. 102, n.º 46, pp. 16569-16572, nov. 2005, doi: 10.1073/pnas.0507655102.

[75] Amazon, «What is Amazon EC2? - Amazon Elastic Compute Cloud». https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/concepts.html (accedido 26 de julio de 2023).

[76] Plotly, «Dash Documentation & User Guide | Plotly». https://dash.plotly.com/ (accedido 26 de julio de 2023).

[77] Ministerio de Ciencia Tecnología e Innovación, «GrupLAC - Grupo de Ingeniería Telemática». https://scienti.minciencias.gov.co/gruplac/jsp/Medicion/graficas/verPerfiles.jsp?id_convocatoria=21&nroIdGrupo=00000000008160 (accedido 26 de julio de 2023).

[78] AnyDesk, «Anydesk», *AnyDesk*. https://anydesk.com/es/caracteristicas (accedido 16 de agosto de 2023).

[79] Amazon, «Instancias T2 de Amazon EC2 – Amazon Web Services (AWS)», *Amazon Web Services, Inc.* https://aws.amazon.com/es/ec2/instance-types/t2/ (accedido 26 de julio de 2023).

[80] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, y A. Wesslén, *Experimentation in Software Engineering*. Berlin, Heidelberg: Springer, 2012. doi: 10.1007/978-3-642-29044-2.

[81] L. Faulkner, «Beyond the five-user assumption: Benefits of increased sample sizes in usability testing», *Behav. Res. Methods Instrum. Comput.*, vol. 35, n.º 3, pp. 379-383, ago. 2003, doi: 10.3758/BF03195514.

ANEXOS

A. ADAPTACIÓN DE LA METODOLOGÍA A PARTIR DE *PMBOK*

Para el desarrollo de este trabajo de grado, se empleó una metodología adaptada de *PMBOK* (*Project Management Body of Knowledge*) o Cuerpo de Conocimiento de Gestión de Proyectos, tomando la edición más actual de la guía *PMBOK* como elemento de referencia [54]. La metodología contempla toda la gama de enfoques de desarrollo y la aplicación de elementos para la gestión de proyectos que perduran en el tiempo y buscan adaptarse adecuadamente a entregables de diferente naturaleza, por lo cual se encuentran en constante evolución y evaluación con el objetivo de cubrir la necesidad de un matiz amplio de funciones que posibilitan nuevas formas de pensar y trabajar en colaboración. Cabe mencionar que, en los últimos 10 años y ante el vertiginoso avance del *software* hacia todo tipo de productos, servicios y soluciones, la transformación del estándar ha sido alineada a dicho ámbito otorgando buenos resultados y referencias según las experiencias de interesados y lideres a nivel mundial. Teniendo en cuenta el contexto académico del presente trabajo de investigación, la naturaleza del problema y la pregunta de investigación, se abordaron los elementos presentados a continuación para adaptar la metodología de desarrollo a seguir en la construcción del proyecto durante sus diferentes etapas:

- **Dominios de desempeño del enfoque de desarrollo y del ciclo de vida:** Aborda las actividades y funciones asociadas con un enfoque de desarrollo consistente con los entregables del proyecto, la cadencia de entrega y el ciclo de vida del proyecto que consta de fases conectadas a la entrega de valor de los interesados de principio a fin.
- **Dominio de desempeño de la planificación:** Se abordan variables para la planificación de cantidades, momentos y frecuencia que definirán el tipo de enfoque de desarrollo a seguir, el alcance de los entregables y los requisitos organizacionales. Se identifican las entregas, estimación, cronograma y presupuesto como variables debido a que pueden estar sujetas a cambios durante todo el proyecto y sugerir decisiones puntuales sobre los elementos mencionados.
- **Dominio de desempeño de la entrega:** Aborda el seguimiento de los objetivos del proyecto, la verificación de los requisitos, la validación por parte de los interesados, y el cumplimiento de las expectativas de calidad para producir los entregables esperados como resultado.
- **Artefactos:** Se aborda la creación de una estructura de desglose del trabajo que permita subdividir los entregables y el trabajo del proyecto en

componentes más pequeños y manejables. Se opta por este artefacto bajo recomendación de la guía y dado su extensivo uso en proyectos *software.*

Una vez identificados los principios y las buenas prácticas a seguir, contempladas en la guía de *PMBOK*, se procede a condensar y definir los métodos necesarios para la construcción de un sistema que permita extraer, persistir, analizar y visualizar datos bibliográficos en el ámbito de los grupos de investigación en el Cauca.

Enfoque de desarrollo

Según [54], el enfoque de desarrollo es el medio utilizado para la creación y desarrollo de un producto, servicio o resultado durante el ciclo de vida del proyecto. Tres de los enfoques de desarrollo más comunes son el predictivo, el hibrido y el adaptativo que a menudo se visualizan como un espectro observado en la Figura anexos 1. El enfoque predictivo es recomendado cuando los requisitos del proyecto pueden definirse, recopilarse y analizarse al comienzo del proyecto, en este enfoque las variables como el alcance, cronograma, costo, recursos y riesgo pueden quedar bien definidas en las fases tempranas del ciclo de vida del proyecto.

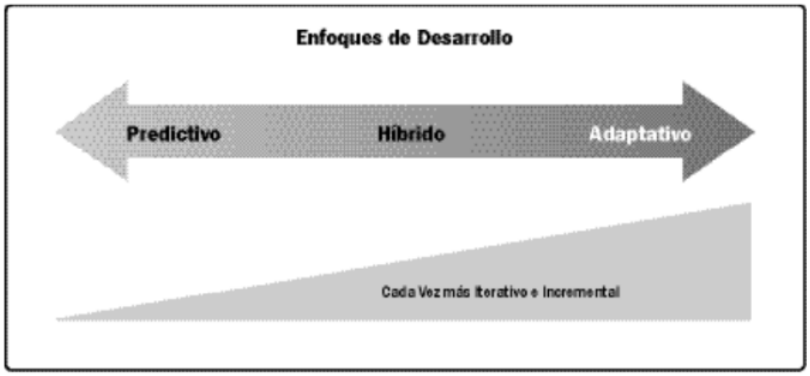


Figura anexos 1. Enfoques de desarrollo. Tomado de [54].

El enfoque adaptativo es recomendado cuando los requisitos están sujetos a un alto nivel de incertidumbre y volatilidad por lo que es probable que cambien a lo largo del proyecto, por lo general tiende a ser iterativo e incremental debido a que evoluciona y se retroalimenta constantemente durante su ciclo de vida.

El enfoque híbrido es una combinación de ambos enfoques por lo que permite el uso de elementos de cada uno, es recomendado cuando existe incertidumbre o riesgo en torno a los requisitos, aunque no de manera critica. También es útil cuando los entregables de un proyecto pueden ser modularizados y desarrollados por partes del equipo, por estas razones se optó por implementar un **enfoque de desarrollo híbrido** dado que el proyecto presenta cierta incertidumbre inicial en sus requisitos

pues actualmente no existen herramientas de análisis bibliométrico adaptadas a las fuentes de datos, niveles de granularidad e interesados para la investigación, además de presentarse cierto grado de volatilidad en la estructura de las fuentes de datos y la posibilidad de cambios debidos a actualizaciones que pueden afectar las variables de planificación. También se destacó el beneficio de modularizar los entregables que engloban los objetivos específicos del presente proyecto.

Modularidad

Como primer paso se procedió a dividir el proyecto en tres módulos distribuyendo las fuentes de datos y las funciones necesarias. Los módulos representan subsistemas o entregables que, una vez integrados, consolidan un sistema completo que presenta características deseadas en la ingeniería de *software*, como un bajo acoplamiento y una alta cohesión. Cuanto más bajo es el acoplamiento menor es la dependencia entre los módulos y por tanto el riesgo de propagación de errores. Cuanta más alta es la cohesión mejor es la medida de relación entre los módulos y por tanto su respuesta objetivo. Los módulos propuestos fueron: **Módulo Extractor *CVLAC-GrupLAC,* Módulo Extractor *Scopus*** y **Módulo *Dashboard***. Cada uno de los módulos propuestos requiere de sus propias fases de desarrollo para su modelamiento, construcción y evaluación.

Enfoque Iterativo-Incremental

El enfoque de desarrollo hibrido permite adaptar y utilizar a su vez un carácter iterativo-incremental para el desarrollo del proyecto. La característica iterativa es útil para aclarar los requisitos de un módulo e investigar diversas opciones entre las fases de la iteración, permitiendo una ventaja adaptativa a lo largo de su implementación. Cada una de las iteraciones cuenta con las fases de desarrollo observadas en la Figura anexos 2. Siguiendo este enfoque, las fases de desarrollo no están estrictamente ligadas a un solo ciclo u ordenamiento como en otras metodologías, sino que permiten la flexibilidad de moverse entre ellas hasta lograr un resultado aceptable a partir de la retroalimentación o hallazgos.

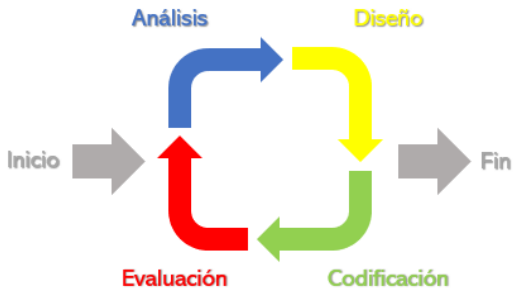


Figura anexos 2. Fases de desarrollo iterativo. Adaptado de [54].

La característica incremental del enfoque permite que cada iteración añada funcionalidad dentro de un marco de tiempo determinado, en este caso cada iteración añade un subsistema idealmente independiente al sistema. El entregable final se considera terminado una vez superada la última iteración. En total se consideran 3 iteraciones incrementales correspondientes a los 3 módulos, contando cada una de ellas con las 4 fases de desarrollo y con la posibilidad de hacer transiciones flexibles entre las fases o la repetición de estas dentro de una misma iteración. De esta forma se consolida un enfoque tanto iterativo como incremental, con características predictivas en sus etapas iniciales y adaptativas durante su transición.

Ciclo de vida

Con base en las consideraciones anteriores se procede a definir el ciclo de vida del proyecto, representado en la Figura 3.1. El ciclo de vida del proyecto es la serie de fases que representan la evolución del producto, es decir, su inicio, crecimiento, madurez y retiro.

Estructura de desglose del trabajo

La Estructura de Desglose del Trabajo o *EDT,* es una descomposición jerárquica del alcance o visión del trabajo a ser realizado por los integrantes del proyecto para lograr los objetivos y crear los entregables requeridos. La *EDT* organiza y define el proyecto representando el trabajo planeado, el cual es contenido en los niveles más bajos de los componentes de la estructura. Estos son llamados paquetes de trabajo y son usados para agrupar las actividades derivadas. Los niveles más altos de los componentes representan a los entregables o módulos del proyecto. La *EDT* del presente trabajo de investigación se observa en la Figura anexos 3, se tienen en cuenta los elementos desarrollados durante este capítulo, las actividades necesarias y la documentación requerida como entregable.

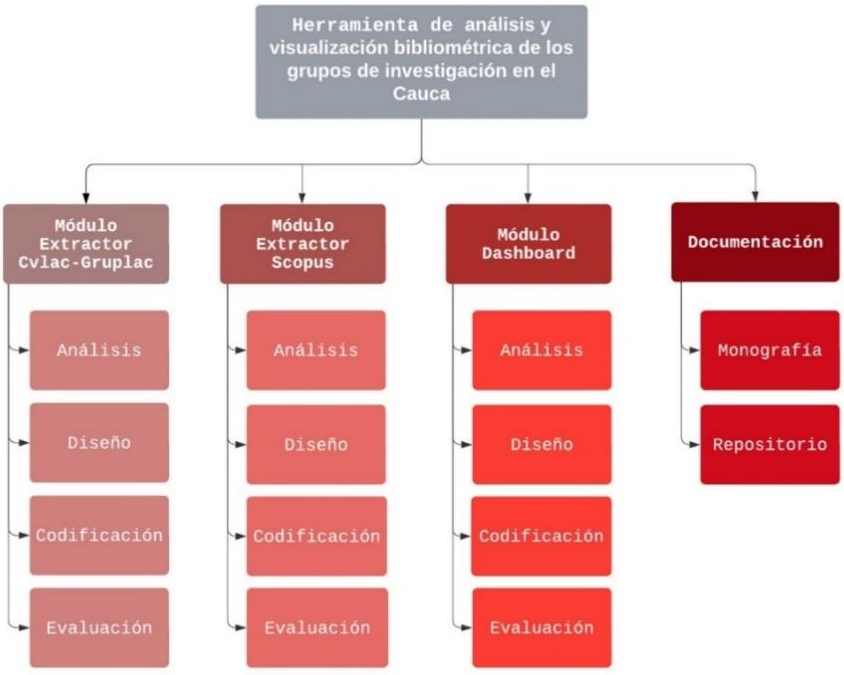


Figura anexos 3. Estructura de Desglose del Trabajo. Adaptado de [54].

Los capítulos 4, 5, 6 y 7 de esta monografía profundizan en el trabajo realizado durante este proyecto de grado, siguiendo la ruta metodológica sugerida por el ciclo de vida del proyecto y la *EDT*, entre otros instrumentos y criterios adaptados del cuerpo de conocimiento de la gestión de proyectos. También se profundiza en cada uno de los paquetes de trabajo o fases de desarrollo y se desglosan sus actividades.

B. RECURSOS Y DOCUMENTACIÓN.

Se pone a disposición un repositorio digital que hospeda el código fuente, recursos y documentación vinculados con el *Dashboard*. El mencionado repositorio se encuentra alojado en la plataforma GitHub bajo la siguiente dirección:

https://github.com/alexper11/General

Este repositorio ofrece una estructura organizada que permite acceder y revisar de manera detallada todos los componentes del proyecto. Contiene el código fuente utilizado en el desarrollo, así como como instrucciones de instalación, imágenes, archivos de configuración, modelos, entre otros elementos que contribuyeron al proceso de implementación.

C. PSEUDOCÓDIGOS DE ALGORITMOS PARA LA EXTRACCIÓN Y PROCESAMIENTO DE LOS DATOS

Algoritmo 1. Extracción de una hoja de vida en CVLAC

Para la extracción de una hoja de vida proveniente de *CVLAC,* se programa la clase "*ExtractorCvlac"* con sus atributos y métodos. En la Tabla anexos 1 se observa el pseudocódigo de un algoritmo para la extracción de una hoja de vida. El símbolo "*#*" es usado como comentario de la línea y las flechas simbolizan la asignación de valores. Los algoritmos únicos mencionados anteriormente, usan el prefijo "*get"* dentro de un objeto de la clase, como por ejemplo "*get_basico*". El funcionamiento interno de estos es omitido debido a su compleja y diversa composición, por lo tanto, su propósito se expresa a partir de sus entradas y salidas. Los métodos únicos también cuentan con la particularidad de actualizar internamente su atributo asociado cuando son ejecutados. Existen métodos no anidados en la clase que son de uso general, como "*get_lxml"*, que sirve para formatear los documentos *web*. La palabra "*dataframe"* hace referencia a la estructura de datos utilizada, la cuál es flexible para su adaptación a formatos de bases de datos o formatos de lectura como *CSV* (*Comma Separated Values)* y permite una adecuada manipulación de los datos. El símbolo del punto "." dentro del pseudocódigo, se usa como medio de acceso a los atributos o métodos de un objeto.

	Algoritmo para la extracción de una hoja de vida.
1:	Crear objeto *E* de clase *ExtractorCvlac* con atributos vacíos
2:	***Método*** *get_cv(url)* de *E* # *"url"* es un parámetro que representa la hoja de vida en cuestión
3:	***Método*** *get_lxml(url)* # Formatear "*DOM*"
4:	Solicitar *DOM* de *url*
5:	*soup* ← contenido formateado de *DOM* # *"soup"* es una variable
6:	Retorna *soup*
7:	***Métodos*** *get_<tabla>(soup,url)* de *E* # Algoritmos únicos para CVLAC
8:	**…** #Encuentran la tabla de información del atributo "*<tabla>"* en el objeto "*soup"*
9:	**…** #Procesan y Almacenan los datos en el atributo "*<tabla>"* de "*E"*
10:	**…** #Concatenan los valores anteriores a los actuales en "*<tabla>*" en caso de iteraciones
11:	**…** #Retornan una estructura de datos organizada y granulada
12:	*dataframe_basico* ← *E.get_basico(soup,url)* # *E.basico == dataframe_basico*
13:	*dataframe_articulos* ← *E.get_articulos(soup,url)* # *E.articulos == dataframe_articulos*
14:	**…** # Tablas restantes
15:	Limpiar atributos de *E* # Opcional
16:	Retornar diccionario de *dataframe_<tabla>* # Contiene *"dataframes"* de cada tabla del perfil

Tabla anexos 1. Pseudocódigo de extracción de un perfil *CVLAC*. Fuente propia.

Algoritmo 2. Extracción del perfil de un grupo de investigación

Para la extracción de un perfil de *GrupLAC* se aplica el algoritmo expuesto en la Tabla anexos 2, que es muy similar al algoritmo anterior. Se utiliza la clase "*ExtractorGruplac"* y los atributos de prefijo "***perfil"*** de esta, así como sus métodos asociados. El algoritmo utilizado sirve para uno o muchos enlaces *web* y no retorna un diccionario como el algoritmo anterior, dado que este se desarrolla con la intención de acumular muchos perfiles.

	Algoritmo para la extracción del perfil de un grupo de investigación.
1:	Crear objeto *EG* de clase *ExtractorGruplac* con atributos vacíos
2:	***Método*** *set_perfil_attrs([url])* de *EG* **#** *"[url]"* es una lista con solo una "*url"*
3:	**Para** cada *url* **en** *[url]* # Única iteración
4:	***Método*** *get_lxml(url)*
5:	Solicitar *DOM* de *url*
6:	s*oup* ← contenido formateado de *DOM*
7:	Retorna *soup*
8:	***Métodos*** *get_perfil_<tabla>(soup,url)* de *EG* # Algoritmos únicos para GrupLAC
9:	**…** #Encuentran la tabla de información del atributo "*perfil_<tabla>"* en el objeto "*soup"*
10:	**…** #Procesan y Almacenan los datos en el atributo "*perfil_<tabla>"* de "*EG"*
11:	*EG.get_perfil_basico(soup,url)* # Actualiza atributo "*perfil_basico*" de *"EG"* con datos de "*url*"
12:	*EG.get_perfil_articulos(soup,url)* #Actualiza atributo "*perfil_articulos*" con datos de "*url*"
13:	**…** # Actualizar atributos de tablas restantes
14:	Acceder a atributos *EG.perfil_<tabla>* **#** "*EG*" contiene los datos organizados en sus atributos

Tabla anexos 2. Pseudocódigo de extracción de un perfil *GrupLAC.* Fuente propia.

Algoritmo 3. Extracción de hojas de vida de los miembros de un grupo de investigación

La clase "*ExtractorGruplac"* también ofrece la opción de extraer las hojas de vida de los integrantes del grupo de investigación a través de los atributos de prefijo "*grup"* que utilizan los métodos de la clase padre, puesto que ya abordan la extracción de una hoja de vida. Sin embargo, estos atributos tienen la particularidad de acumularlas con base en la lista que se encuentra en la tabla de integrantes de un perfil de grupo. Esta acumulación se da por la iteración de los métodos que permiten concatenar los resultados en estos atributos. En la Tabla anexos 3 se encuentra el algoritmo para hacer la extracción de un grupo de hojas de vida perteneciente a un perfil de *GrupLAC*.

Algoritmo para la extracción de un conjunto de hojas de vida pertenecientes a un grupo de investigación.	
1:	Crear objeto *EG* de clase *ExtractorGruplac* con atributos vacíos
2:	***Método*** *get_cvs(url_gruplac)* de *EG*
3:	***Método*** *get_members_list(url_gruplac)* de *EG*
4:	Crear variable vacia *lista_m* para la lista de miembros
5:	Solicitar *DOM* de *url_gruplac*
6:	s*oup* ← contenido formateado de *DOM*
7:	Listar en *url_inv* todos los enlaces presentes en *soup*
8:	**Para** cada *url* **en** *url_inv*
9:	**Si** *url* **contiene** "*… scienti.minciencias.gov.co/cvlac…*" **entonces** *lista_m* agrega *url*
10:	Retorna *lista_m*
11:	**Para** cada *m* **en** *lista_m*
12:	*soup* ← *get_lxml(m)* # Obtener *DOM* formateado
13:	*dataframe_basico* ← *EG.get_basico(soup,m)* # "*dataframe_basico"* guarda "*basico*" de "*EG*"
14:	**…** #internamente "*basico*" concatena su extracción actual con las anteriores en cada ciclo
15:	*dataframe_articulos* ← *EG.get_articulos(soup,m)*
16:	**…** #internamente "articulos" concatena su extracción actual con las anteriores en cada ciclo
17:	**…** # Tablas restantes, guardar todos los "*dataframe_<tabla>"*
18:	Limpiar atributos de *EG* # Opcional
19:	Retornar diccionario de *dataframe_<tabla>* # Contiene todas las hojas de vida del grupo
20:	*EG.grup_<tabla>* ← *dataframe_<tabla>*
21:	**…** # Asignación por cada "*<tabla>*"
22:	Acceder a atributos *EG.grup_<tabla>*

Tabla anexos 3. Pseudocódigo de extracción de conjunto de hojas de vida pertenecientes a un perfil *GrupLAC.* Fuente propia.

Algoritmos 4 y 5. Operaciones sobre las bases de datos generadas

En la Tabla anexos 4 se presenta el pseudocódigo para el algoritmo de inserción de datos en las tablas de las bases de datos. Este algoritmo está presente en todos los controladores del subsistema con variaciones en la estructura de datos especifica que manejan. El algoritmo de la Tabla anexos 5 presenta las mismas condiciones y se encarga de eliminar los registros de una tabla en la base de datos de *CVLAC* asociados a un código "*idcvlac"*. El mismo algoritmo se emplea para *GrupLAC* con variaciones del nombre del método, el código "*idgruplac*" y el modelo a usar. La combinación de estos algoritmos, en sus diferentes controladores, permiten consultar la presencia de un perfil especifico en las bases de datos para actualizarlo mediante el borrado de datos anteriores y la inserción de datos extraídos en el instante. Si bien se usan por separado, es posible hacer la inserción masiva de datos o el borrado permanente.

Algoritmo para la inserción de datos extraídos en las bases de datos.	
1:	Crear objeto *MC* de clase *<modelo>Controller # ej. "ArticulosController"* usa el modelo "*Artículos"*
2:	***Método*** *insert_df(dataframe)* de *MC* # parámetro "*dataframe"* es la tabla extraída de internet
3:	*dicList* ← Reformatear *dataframe*
4:	Levantar sesión en <*database*> # "<*database*>*"* es "c*vlacdb*" o "*gruplacdb*"
5:	***Método*** *bulk_insert_mappings(<modelo>,dicList)* # Función incorporada, insertar datos
6:	***Método*** *commit()* # Función incorporada, comprometer inserción en "<*database*>*"*
7:	***Excepción*** *rollback()* # Función incorporada, no efectuar inserción en "<*database*>*"*
8:	Cerrar sesión en <*database*>

Tabla anexos 4. Pseudocódigo de algoritmo de inserción en base de datos. Fuente propia.

Algoritmo para el borrado de datos presentes en las bases de datos.	
1:	Crear objeto *MC* de clase *<modelo>Controller # ej. "BasicoController"* usa el modelo "*Basico"*
2:	***Método*** *delete_idcvlac(idcvlac)* de *MC* # "***delete_idgruplac(idgruplac)"*** es el algoritmo análogo
3:	Levantar sesión en *cvlacdb* # "*gruplacdb*" es la base de datos análoga
4:	***Método*** *query(<modelo>)* # Función incorporada, consultar tabla
5:	***Método*** *filter(<modelo>.idcvlac==idcvlac)* # Función incorporada, filtrar registros
6:	***Método*** *delete()* # Función incorporada, eliminar registros
7:	***Método*** *commit()* # Efectuar borrado en "*cvlacdb"*
8:	***Excepción*** *rollback()* # No efectuar borrado
9:	Cerrar sesión en *cvlacdb*

Tabla anexos 5. Pseudocódigo de algoritmo de borrado en base de datos. Fuente propia.

Algoritmo 6. Extracción masiva en *CVLAC* Y *GrupLAC*

La extracción masiva de datos se realiza a partir de la implementación de todos los métodos desarrollados. Siguiendo lo mencionado anteriormente en el Algoritmo 2, se aprovecha la iteración de enlaces *web* para recorrer todos los grupos de investigación reconocidos por Minciencias en el Departamento del Cauca según lo planteado en la Fase de Análisis del módulo. En la Tabla anexos 6 se observa el pseudocódigo del algoritmo utilizado para la extracción masiva y persistencia de los datos.

Algoritmo para la extracción y persistencia de la información de los perfiles de investigación a nivel regional	
1:	Crear base de datos *cvlacdb* con todas sus tablas
2:	**Si** *cvlacdb* existe **entonces** limpia todas las tablas de la base de datos
3:	Crear base de datos *gruplacdb*
4:	**Si** *gruplacdb* existe **entonces** limpia todas las tablas de la base de datos
5:	Crear objeto *EG* de clase *ExtractorGruplac*
6:	***Método*** *get_gruplac_list(url)* de *EG* # El enlace en "*url*" contiene los GrupLAC a nivel Cauca

7:	Crear lista vacia *gruplac_list*
8:	Solicitar *DOM* de *url*
9:	s*oup* ← contenido formateado de *DOM*
10:	Listar en *links* todos los enlaces presentes en *soup*
11:	**Para** cada *link* **en** *links*
12:	**Si** *link* **contiene** "*…scienti.minciencias.gov.co/gruplac…"* **entonces** *gruplac_list* agrega *link*
13:	Retornar *gruplac_list*
14:	*gruplac_list* ← *EG.get_gruplac_list(url)* #Lista de GrupLAC a nivel Cauca
15:	***Método*** *set_grup_attrs(gruplac_list)* de *EG* # Almacena y acumula las hojas de vida de cada grupo
16:	**Para** cada *gruplac* **en** *gruplac_list*
17:	*dataframes* ← *EG.get_cvs(gruplac)* # "*dataframes"* contiene todos los "*dataframe_<tabla>"*
18:	**…** # Algoritmo en Tabla anexos 3
19:	*EG.grup_basico* ← *EG.grup_basico* **concatena** *dataframe_basico* #Actualizar por iteración
20:	*EG.grup_articulos* ← *EG.grup_articulos* **concatena** *dataframe_basico*
21:	**…** # Actualizar atributos "*grup_<tabla>*" de tablas restantes
22:	Acceder a atributos *EG.grup_<tabla>* # Contienen todos los CVLAC de la lista de GrupLAC
23:	***Extracción masiva y persistencia en CVLAC***
24:	*EG.set_grup_attrs(gruplac_list)*
25:	Crear objeto *BC* de clase *BasicoController* # Objeto controlador de CVLAC
26:	*BC.insert_df(EG.grup_basico)* # Insertar en modelo "*Basico"* el contenido de "*grup_basico*"
27:	Crear objeto *AC* de clase *ArticulosController*
28:	*AC.insert_df(EG.grup_articulos)* # Insertar en modelo "*Articulos"* el contenido de "*grup_articulos*"
29:	**…** # Repetir el proceso con los controladores y atributos de las tablas de CVLAC restantes
30:	***Extracción masiva y persistencia en GrupLAC***
31:	*EG.set_perfil_attrs(gruplac_list)* **#** Rellenar atributos "*perfil_<tabla>*" de "*EG*"
32:	**…** # Algoritmo en Tabla anexos 2
33:	Crear objeto *BGC* de clase *BasicoGController* # Objeto controlador de GrupLAC
34:	*BGC.insert_df(EG.perfil_basico)* # Insertar en modelo "*BasicoG"* el contenido de "*perfil_basico*"
35:	Crear objeto *AGC* de clase *ArticulosGController*
36:	*AGC.insert_df(EG.perfil_articulos)* # Insertar en modelo "*ArticulosG"* el contenido de "*perfil_articulos*"
37:	**…** # Repetir el proceso con los controladores y atributos de las tablas de GrupLAC restantes
38:	Reportar Extracción finalizada

Tabla anexos 6. Pseudocódigo de algoritmo de extracción masiva y persistencia de datos. Fuente propia.

Algoritmo 7. Extracción de los datos de un autor en *Scopus*

La Tabla anexos 7 contiene el algoritmo para la extracción de un autor en *Scopus*.

Algoritmo para la extracción de datos de un autor especifico de Scopus	
1:	Crear objeto *ES* de clase *ExtractorScopus*

2:	**Método** *get_authors_df([author])* de *ES* # parámetro "***[author]"*** es una lista de un solo código de autor
3:	***API Author Retrieval*** Solicitar datos de *author* # Única iteración
4:	*result* ← Respuesta JSON retornada
5:	*ES.autores* ← Extraer campos de *result* # "*autores*" es un atributo de "*ES"*
6:	**… #** Métodos de extracción de características para rellenar el diccionario *autores*
7:	*df_autores* ← *ES.autores*
8:	Retornar *df_autores* # "*df_autores*" contiene una tabla con la información del autor

Tabla anexos 7. Pseudocódigo de algoritmo de extracción de un autor. Fuente propia.

Algoritmo 8. Extracción de los datos de un producto en *Scopus*

La extracción de un producto se da mediante el algoritmo de la Tabla anexos 8. Los algoritmos están programados para actualizar los datos de los atributos de la clase extractora. Algunos métodos de este módulo están adaptados para recibir un único elemento identificador y otros para recibir listas de estos elementos.

Algoritmo para la extracción de datos de un producto especifico de Scopus	
1:	Crear objeto *ES* de clase *ExtractorScopus*
2:	**Método** *get_article(article)* de *ES* # parámetro "***article"*** es un código de producto
3:	***API Abstract Retrieval*** Solicitar datos de *article*
4:	*result* ← Respuesta JSON retornada
5:	*ES.productos* ← Extraer campos de *result* # "*productos*" es un atributo de "*ES"*
6:	**… #** Métodos de extracción de características para rellenar el diccionario *productos*
7:	*df_productos* ← *ES.productos*
8:	Retornar *df_autores*

Tabla anexos 8. Pseudocódigo de algoritmo de extracción de un producto. Fuente propia.

Algoritmo 9. Extracción masiva y persistencia de datos de *Scopus*

Los controladores del módulo *Scopus* son capaces de hacer inserciones de estructuras de datos extraídas desde la fuente hacía la base de datos propia mediante la función *"insert_df"* presente en los dos controladores. El proceso es similar al presentado en la Tabla anexos 6, pero en este caso se tiene en cuenta la base de datos de este módulo llamada "*scopusdb*" y las tablas de autores y productos.

Las funciones de los controladores *"delete_affil_id"* y *"delete_autor_id",* están adaptadas de los procesos de la Tabla anexo 5, teniendo en cuenta el remplazo de los códigos "*idcvlac"* usados en ella, a códigos de afiliación y de autor de *Scopus,* como sugieren los nombres de estos dos métodos. Otra diferencia se encuentra en el paso 5 del algoritmo donde, en lugar de buscar un código exacto con la instrucción "==", se busca que el código este contenido en el dato de la celda pues este dato

podría estar compuesto por diversos códigos concatenados y relacionados al mismo elemento en cuestión. Los datos de códigos concatenados y distinguidos por separadores de punto y coma pueden ser de afiliaciones, productos o autores registrados para un mismo elemento. Luego los dos métodos de borrado permiten modificar las tablas de "*scopusdb*" con base en los códigos de afiliación o autor contenidos en las columnas que albergan estas características de los registros.

La extracción masiva de datos a nivel del Departamento del Cauca necesita utilizar todos los métodos de las clases desarrolladas para este módulo en sus etapas de extracción y persistencia de datos. La Tabla anexos 9 contiene el algoritmo de la extracción a nivel regional, haciendo uso de la lista de afiliaciones generada a partir del análisis de la recolección de datos.

	***Algoritmo para la extracción de datos de Scopus* a nivel del Departamento del Cauca**
1:	Crear objeto *ES* de clase *ExtractorScopus*
2:	Crear objeto *AC* de clase *AutoresController*
3:	Crear objeto *PC* de clase *ProductosController*
4:	Definir *cauca_affiliations* # Variable de lista de afiliaciones en el Cauca
5:	Definir *authors_set* # Variable para conjunto de autores
6:	**Para** cada *affiliation* **en** *cauca_affiliations*
7:	*authors_set* agrega *ES.get_auid_list(affiliation)* # Conjunto de códigos de autor
	… ***API Author Search*** Solicitar autores de *affiliation*
8:	*dataframe_authors* ← *ES.get_authors_df(authors_set)* # Tabla 3, rellenar datos de autores
9:	*AC.insert_df(dataframe_authors)* # Insertar en tabla de autores de "*scopusdb*"
10:	**Método** *get_articles_full(cauca_affiliations)* de *ES*
11:	**Para** cada *affiliation* en *cauca_affiliations*
12:	*articles* ← *ES.get_eid_list(affiliation)* # Lista de códigos de producto en variable "*articles"*
13:	**…** ***API Scopus Search*** Solicitar productos de *affiliation*
14:	**Para** cada *article* en *articles*
15:	**…** Rellenar datos de *ES.productos* # Similar a Tabla 4
16:	*dataframe_products* ← *ES.productos*
17:	Retornar *dataframe_products*
18:	*dataframe_products* ← *ES.get_articles_full(cauca_affiliations)*
19:	*PC.insert_df(dataframe_products)* # Insertar en table de productos de "*scopusdb*"
20:	Reportar extracción finalizada

Tabla anexos 9. Pseudocódigo de algoritmo de extracción masiva y persistencia de datos. Fuente propia.

Algoritmo 10. Integración de datos entre GrupLAC y *Scopus*

En la Tabla anexos 10 se observa el pseudocódigo del algoritmo desarrollado para la integración de datos entre *GrupLAC* y *Scopus*, con el fin de distinguir a que grupos de investigación pertenecen ciertos autores y productos:

	Algoritmo para la integración de datos de los módulos extractores.
1:	**Integración de Autores**
2:	Generar *idcvlac_authorid* # Estructura de datos de los "*idcvlac"* y sus "*authorid*" de *Scopus*
3:	Generar *idcvlac_orcid* # Estructura de datos de los "*idcvlac"* y sus "*orcid*"
4:	Generar *idgruplac_idcvlac* # Estructura de datos de los "*idgruplac*" y sus "*idcvlac*"
5:	Generar *idgruplac_grupo*
6:	*idgruplac_authorid* ← Unir *idcvlac_authorid* e *idgruplac_idcvlac* Por *idcvlac*
7:	*authorid_grupo* ← Unir *idgruplac_authorid* e *idgruplac_grupo* Por *idgruplac*
8:	*idgruplac_orcid* ← Unir *idcvlac_orcid* e *idgruplac_idcvlac* Por *idcvlac*
9:	*orcid_grupo* ← Unir *idgruplac_orcid* e *idgruplac_grupo* Por *idgruplac*
10:	*AutoresRef1* ← Agrupar *authorid_grupo* Por *authorid*
11:	Concatenar características agrupadas # De varios a un solo registro para un mismo *authorid*
12:	*AutoresRef2* ← Agrupar *orcid_grupo* Por *orcid*
13:	Concatenar características agrupadas # De varios a un solo registro para un mismo *orcid*
14:	*Autores1* ← Unir *Autores* y *AutoresRef1* Por *authorid* # "*Autores*" es la tabla original
15:	*Autores2* ← Unir *Autores* y *AutoresRef2* Por *orcid*
16:	*Autores* ← Concatenar *Autores1* y *Autores2* # Apilar dos estructuras de datos
17:	Remover registros duplicados y nulos con base a *idgruplac, idcvlac* y *authorid*
18:	**Integración de Productos**
19:	Observar categorías de la tabla *Productos* de *scopusdb*
20:	Generar *art_scopus* con categorías *Article,Review,Letter,Note,Erratum,Data Paper* y *Short Survey*
21:	Generar *lib_scopus* con categorías *Book* y *Book Chapter*
22:	Generar *otros_scopus* con categorías *Conference Paper* y *Editorial*
23:	**Método** *match_articulos_doi(art_scopus,art_gruplac)* #"*art_gruplac*" es la tabla de artículos *gruplac*
24:	*matched* ← Unir *art_scopus* y *art_gruplac* por *doi* #"*doi*" es la columna con el identificador
25:	*matched* ← Unir *matched* e *idgruplac_grupo*
26:	*matched* ← Agrupar *matched* por *doi*
27:	Concatenar características agrupadas
28:	*art_scopus* ← Unir *art_scopus* y *matched* por *doi*
29:	Retornar *art_scopus* # El anterior es un proceso de emparejamiento de datos
30:	**Método** *match_articulos_nombre(art_scopus,art_gruplac)*
31:	Emparejar *art_scopus* y *art_gruplac* por *nombre* # Incluye procesos de unión, agrupación, etc.
32:	**…** # Estos procesos incluyen normalización, "RegEx" y más tratamientos de cadenas de texto
33:	Retornar *art_scopus*
34:	**Método** *match_libros_isbn(lib_scopus,lib_gruplac)*

35:	Emparejar *lib_scopus* y *lib_gruplac* por *isbn*
36:	**…** # Procesos de unión,agrupación,indización,iteración,entre otros.
37:	Retornar *lib_scopus*
38:	**Método** *match_libros_nombre(lib_scopus,lib_gruplac)*
39:	Emparejar *lib_scopus* y *lib_gruplac* por *nombre*
40:	**…** # Incluye los procesos de emparejamiento anteriores y tratamientos de cadenas de texto
41:	Retornar *lib_scopus*
42:	*result_art* ← *match_articulos_doi(art_scopus,art_gruplac)*
43:	*result_art* ← *match_articulos_nombre(result_art,art_gruplac)*
44:	*result_art* ← *match_articulos_nombre(result_art,otrosart_gruplac)* #Tabla de otros artículos *gruplac*
45:	*result_lib* ← *match_libros_isbn(lib_scopus,lib_gruplac)*
46:	*result_lib* ← *match_libros_isbn(result_lib,otroslib_gruplac)* #Tabla de otros libros *gruplac*
47:	*result_lib* ← *match_libros_isbn(result_lib,caplib_gruplac)* #Tabla de capítulos de libros *gruplac*
48:	*result_lib* ← *match_libros_nombre(result_lib,lib_gruplac)*
49:	*result_lib* ← *match_libros_nombre(result_lib,otroslib_gruplac)*
50:	*result_lib* ← *match_libros_nombre(result_lib,caplib_gruplac)*
51:	*result_otros* ← *match_articulos_doi(otros_scopus,art_gruplac)*
52:	*result_otros* ← **…** # Se ejecutan todos los métodos de prefijo *match*, emparejar otras categorías
53:	*Productos* ← Unir *result_art, result_lib* y *result_otros* por *scopus_id*
54:	Remover registros duplicados y nulos con base a *idgruplac* y *scopus_id*

Tabla anexos 10*.* Pseudocódigo de algoritmo de integración de datos. Fuente propia.

Algoritmo 11. Preprocesamiento de datos

Durante la etapa de limpieza de datos se llevan a cabo procesos de preparación de los datos con el fin de ejecutar las funciones del *dashboard* de manera adecuada. Estos procesos de limpieza dependen del tipo de característica, por lo que se implementaron múltiples algoritmos específicos correspondientes a cada combinación tabla – columna que fue seleccionada para ser explotada por el *dashboard*. Como resultado se codificó una cantidad elevada de procesos que son resumidos en la Tabla anexos 11 aludiendo a los tipos de datos posibles que fueron considerados para la totalidad de características diferentes y pertenecientes a todas las tablas de las bases de datos generadas. Cabe mencionar que se construyeron múltiples diccionarios en el código para la normalización de datos categóricos con tal de reforzar la limpieza de editoriales, revistas y afiliaciones en general. Los procesos presentados a continuación fueron aplicados, en un inicio, a todas las características y posteriormente a cada tipo de dato según sus requerimientos puntuales para cada una de las tablas de información.

Algoritmos para la etapa de preprocesamiento de datos del dashboard	
1:	Recuperar conjuntos de datos # Desde bases de datos o archivos locales
2:	Rellenar valores nulos con cadena “No Aplica”

3:	Estandarizar separadores de texto de múltiples valores con ";"
4:	Formatear características con valores de texto, categorías, números, fechas o booleanos
5:	Eliminar espacios en blanco y saltos de línea irrelevantes # iniciales, finales, seguidos, etc.
6:	Generar diccionarios personalizados para datos categóricos # pareja de datos: normalizado-variación
7:	**Variables categóricas**
8:	Eliminar símbolos no alfanuméricos # Uso de *regex* personalizado en cada caso
9:	Definir etiquetas de categoría
10:	Normalizar etiquetas únicas por diversas técnicas
11:	Codificar texto en "*ascii*"
12:	Decodificar texto en "utf-8"
13:	Eliminar acentos, mayúsculas, caracteres residuales, entre otros
14:	Agrupar datos por etiquetas únicas
15:	Unificar coincidencias de registros
16:	Remplazar texto de etiquetas finales por el texto original de mayor repetición
17:	Buscar categoría más repetida por su identificador
18:	Recuperar texto original
19:	Normalizar etiquetas por diccionarios personalizados
20:	**Variables de texto**
21:	Eliminar caracteres irrelevantes # *Regex* personalizado en cada caso, ej: triple comilla, html, etc
22:	Verificar integridad y formato de las cadenas de texto # Espacios, saltos, nulos, escapes, etc
23:	**Variables numéricas**
24:	Eliminar valores irrelevantes # Inspección en cada caso, ej: citas negativas ,*index* decimal, etc.
25:	Verificar valores atípicos
26:	Verificar integridad y formato de registros numéricos # Garantizar operabilidad de datos
27:	**Variables de fecha**
28:	Definir periodo de la característica
29:	Verificar integridad y formato de registros # Garantizar operabilidad de datos
30:	**Variables *booleanas***
31:	Rastrear inconsistencias # Garantizar existencia de dos tipos de valores únicos
32:	Eliminar inconsistencias # incluye eliminar valores vacíos, residuales e irrelevantes según el caso
33:	Rastrear registros duplicados
34:	Unificar registros según coincidencias múltiples # nombres, *DOI, ISBN,* revista, fecha, institución, títulos, editorial, autores, códigos e identificadores según el caso particular de cada tabla
35:	Generar estadísticas de interés de los conjuntos de datos finales
36:	Guardar conjuntos de datos preprocesados

Tabla anexos 11. Pseudocódigo de algoritmos de preprocesamiento de datos del *Dashboard*. Fuente propia.

D. SELECCIÓN DE AFILIACIONES DEL DEPARTAMENTO DEL CAUCA EN *SCOPUS*

Para llevar a cabo la selección de afiliaciones en el Departamento del Cauca se utilizó la *API* de *Scopus* llamada *Affiliation Search API*. Se procedió a realizar consultas individuales utilizando el nombre de cada municipio en el departamento como palabras clave, con el objetivo de obtener todos los posibles resultados para cada región. Se incluyó la capital y el nombre propio del departamento en las búsquedas para compensar excepciones en las que los usuarios digitan una ciudad o país erróneamente.

Posteriormente, se llevó a cabo una minuciosa revisión y análisis de cada afiliación para asegurar que esta perteneciera al Departamento del Cauca a partir de la información disponible en sus sitios *web* oficiales. Aquellas afiliaciones que resultaron no pertenecer a la región del Cauca fueron excluidas del análisis.

Finalmente, se realizó una revisión para eliminar perfiles de afiliaciones sin registros de documentos o autores. Algunas firmas de afiliación contaban con múltiples perfiles conteniendo registros de actividad científica, por lo cual se incluyeron con la intención de unificar estos perfiles en posteriores etapas de limpieza de datos. El proceso de selección de afiliaciones del Cauca desde la base de datos de *Scopus* se presenta resumido en la Tabla de anexos 12.

Región	Cadena de Consulta	Número de Resultados Obtenidos	Número de Resultados Escogidos
Santander de Quilichao	*https://api.elsevier.com/content/search/affiliation ?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("santander de quilichao")) OR (AFFILCITY("santander de quilichao"))))&start=0&count=200*	1	1
Cajibío	*https://api.elsevier.com/content/search/affiliation ?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("cajibio")) OR (AFFILCITY("cajibio"))))&start=0&count=200*	2	2
Patía	*https://api.elsevier.com/content/search/affiliation ?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("patia")) OR (AFFILCITY("patia"))))&start=0&count=200*	1	1
Caldono	*https://api.elsevier.com/content/search/affiliation ?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("caldono")) OR (AFFILCITY("caldono"))))&start=0&count=200*	1	1

Guapi	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("guapi")) OR (AFFILCITY("guapi"))))&start=0&count=200*	1	1
Puracé	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("purace")) OR (AFFILCITY("purace"))))&start=0&count=200*	4	4
San Sebastián	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("san sebastian")) OR (AFFILCITY("san sebastian"))))&start=0&count=200*	5	1
Buenos Aires	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("buenos aires")) OR (AFFILCITY("buenos aires"))))&start=0&count=200*	8	0
Caloto	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("caloto")) OR (AFFILCITY("caloto"))))&start=0&count=200*	0	0
Corinto	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("corinto")) OR (AFFILCITY("corinto"))))&start=0&count=200*	0	0
Guachené	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("guachene")) OR (AFFILCITY("guachene"))))&start=0&count=200*	0	0
Miranda	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("miranda")) OR (AFFILCITY("miranda"))))&start=0&count=200*	0	0
Padilla	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("padilla")) OR (AFFILCITY("padilla"))))&start=0&count=200*	14	0
Puerto Tejada	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("puerto tejada")) OR (AFFILCITY("puerto tejada"))))&start=0&count=200*	0	0
Suarez	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("suarez")) OR (AFFILCITY("suarez"))))&start=0&count=200*	13	0
Villa Rica	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("villa rica")) OR (AFFILCITY("villa rica"))))&start=0&count=200*	0	0

El Tambo	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("el tambo")) OR (AFFILCITY("el tambo"))))&start=0&count=200*	0	0
La Sierra	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("la sierra")) OR (AFFILCITY("la sierra"))))&start=0&count=200*	3	0
Morales	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("morales")) OR (AFFILCITY("morales"))))&start=0&count=200*	3	0
Piendamó	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("piendamo")) OR (AFFILCITY("piendamo"))))&start=0&count=200*	0	0
Rosas	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("rosas")) OR (AFFILCITY("rosas"))))&start=0&count=200*	0	0
Sotará	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("sotara")) OR (AFFILCITY("sotara"))))&start=0&count=200*	0	0
Timbío	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("timbio")) OR (AFFILCITY("timbio"))))&start=0&count=200*	0	0
Almaguer	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("almaguer")) OR (AFFILCITY("almaguer"))))&start=0&count=200*	0	0
Argelia	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("argelia")) OR (AFFILCITY("argelia"))))&start=0&count=200*	1	0
Balboa	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("balboa")) OR (AFFILCITY("balboa"))))&start=0&count=200*	0	0
Bolívar	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("bolivar")) OR (AFFILCITY("bolivar"))))&start=0&count=200*	67	0
Florencia	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("florencia")) OR (AFFILCITY("florencia"))))&start=0&count=200*	40	0

La Vega	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("la vega")) OR (AFFILCITY("la vega"))))&start=0&count=200*	0	0
Mercaderes	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("mercaderes")) OR (AFFILCITY("mercaderes"))))&start=0&count=200*	0	0
Piamonte	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("piamonte")) OR (AFFILCITY("piamonte"))))&start=0&count=200*	1	0
Santa Rosa	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("santa rosa")) OR (AFFILCITY("santa rosa"))))&start=0&count=200*	31	0
Sucre	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("sucre")) OR (AFFILCITY("sucre"))))&start=0&count=200*	31	0
López De Micay	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("lopez de micay")) OR (AFFILCITY("lopez de micay"))))&start=0&count=200*	0	0
Timbiquí	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("timbiqui")) OR (AFFILCITY("timbiqui"))))&start=0&count=200*	0	0
Inzá	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("inza")) OR (AFFILCITY("inza"))))&start=0&count=200*	2	1
Jambaló	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("jambalo")) OR (AFFILCITY("jambalo"))))&start=0&count=200*	0	0
Páez	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("paez")) OR (AFFILCITY("paez"))))&start=0&count=200*	0	0
Silvia	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("silvia")) OR (AFFILCITY("silvia"))))&start=0&count=200*	0	0
Toribio	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("toribio")) OR (AFFILCITY("toribio"))))&start=0&count=200*	0	0

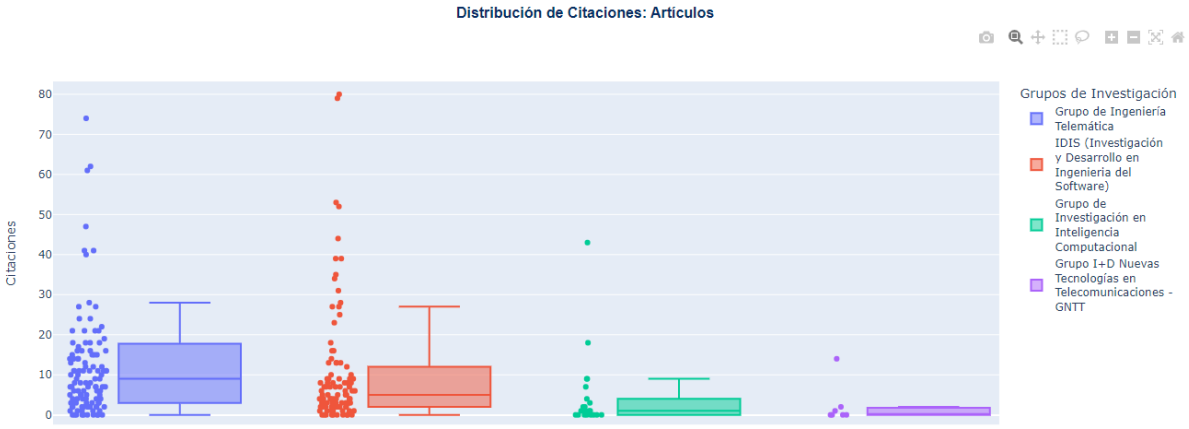
Totoró	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND ((AFFIL("totoro")) OR (AFFILCITY("totoro"))))&start=0&count=200*	0	0
Popayán	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCOUNTRY(colombia)) AND (AFFILCITY(popayan)))&start=0&count=200*	141	109
Cauca	*https://api.elsevier.com/content/search/affiliation?query=((AFFILCITY(cauca)) OR (AFFILCOUNTRY(cauca)))&start=0&count=200*	141	36
TOTAL DE AFILIACONES SELECCIONADAS		131	

Tabla anexos 12. Selección de afiliaciones en el Cauca en *Scopus*. Fuente propia.

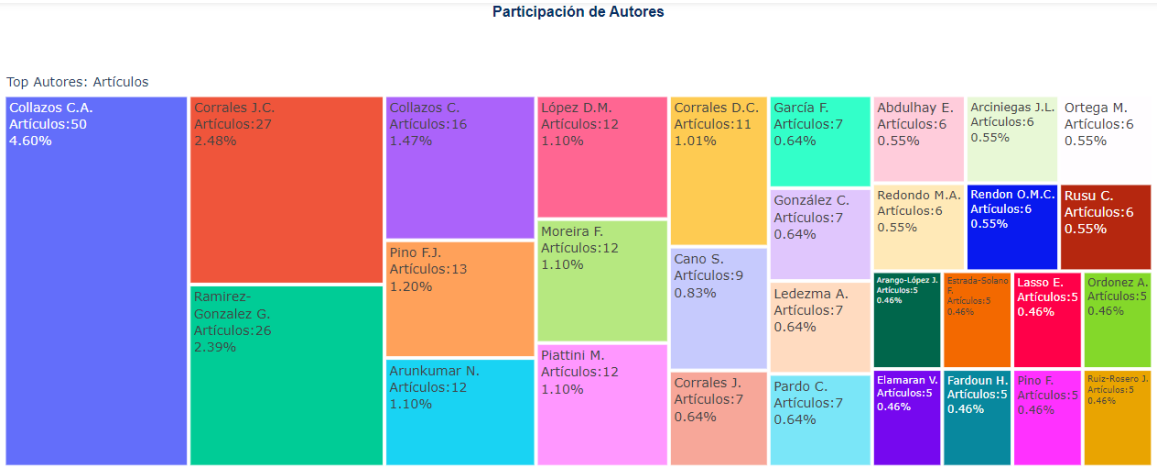
E. OTRAS GRÁFICAS GENERADAS POR EL *DASHBOARD*.

La Figura anexos 4 presenta múltiples gráficas generadas por la herramienta desarrollada en el presente trabajo de investigación. Las gráficas corresponden a diferentes análisis ejecutados teniendo en cuenta distintos criterios de filtrado con el fin de ejemplificar la adaptación que lleva a cabo el procesamiento de los datos en sus cálculos y representación visual. Es de resaltar que la variedad de gráficas que pueden ser generadas es grande y las siguientes muestras sólo buscan compartir con el lector algunos de los resultados posibles. Adicionalmente se ofrece un ejemplo de usabilidad de la herramienta en: https://youtu.be/tdp8FPcWMco

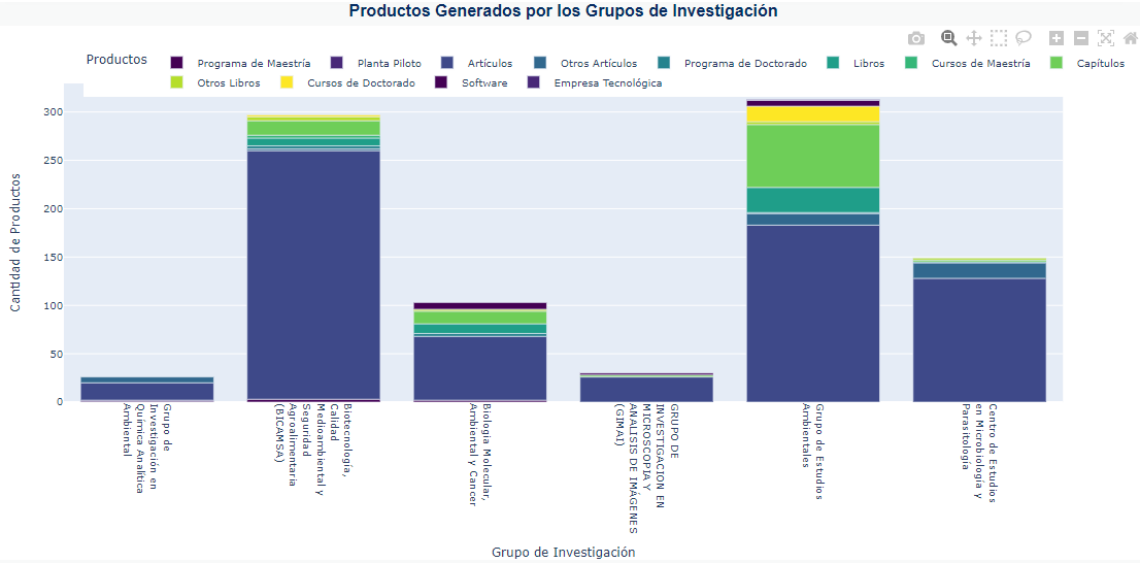
(a)



(b)



(c)



(d)

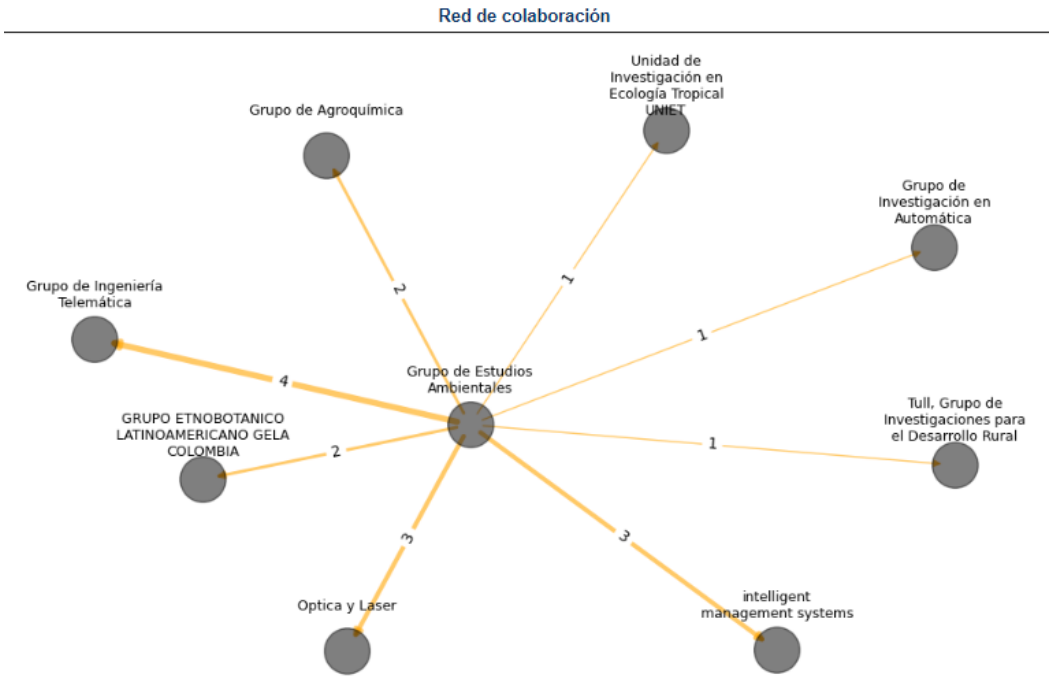


Figura anexos 4. Más ejemplos de gráficas generadas por el *Dashboard*: Diagrama de cajas **(a)**, *treemap* **(b)**, gráfico de barras **(c)** y red de colaboración **(d)**. Fuente propia.

F. EVALUACIÓN DEL PROTOTIPO: *DASHBOARD* PARA EL ANÁLISIS Y VISUALIZACIÓN BIBLIOMÉTRICA DE LOS GRUPOS DE INVESTIGACIÓN EN EL CAUCA.

Preguntas del cuestionario de evaluación

Preguntas	Respuestas
¿Cuál es tu área de trabajo?	Industria
	Academia
	Sector Social
¿Cuál es tu rol principal?	Investigador
	Administrador
	Profesional
	Estudiante
	Otro: _____
La apariencia del *Dashboard* es:	Totalmente agradable
	Moderadamente agradable
	Ni agradable ni desagradable
	Moderadamente desagradable
	Totalmente desagradable
El manejo del *Dashboard* es:	Totalmente fácil
	Moderadamente fácil
	Ni fácil ni complicado
	Moderadamente complicado
	Totalmente complicado
La organización de la información en las ventanas es:	Totalmente clara
	Moderadamente clara
	Ni clara ni confusa
	Moderadamente confusa
	Totalmente confusa
La información gráfica en las secciones del *Dashboard* es:	Totalmente fácil de interpretar
	Moderadamente fácil de interpretar
	Ni fácil ni difícil de interpretar
	Moderadamente difícil de interpretar
	Totalmente difícil de interpretar
La información textual en las secciones del *Dashboard* es:	Totalmente fácil de leer
	Moderadamente fácil de leer
	Ni fácil ni difícil de leer
	Moderadamente difícil de leer
	Totalmente difícil de leer
Las funciones del *Dashboard* son:	Totalmente coherentes

	Moderadamente coherentes
	Ni coherentes ni incoherentes
	Moderadamente incoherentes
	Totalmente incoherentes
El proceso de filtrado de la información es:	Totalmente consistente
	Moderadamente consistente
	Ni consistente ni inconsistente
	Moderadamente inconsistente
	Totalmente inconsistente
Los indicadores bibliométricos utilizados son:	Totalmente relevantes
	Moderadamente relevantes
	Ni relevantes ni irrelevantes
	Moderadamente irrelevantes
	Totalmente irrelevantes
El lenguaje utilizado en el *Dashboard* es:	Totalmente adecuado
	Moderadamente adecuado
	Ni adecuado ni inadecuado
	Moderadamente inadecuado
	Totalmente inadecuado
Podría usar el *Dashboard* sin instrucciones:	Totalmente de acuerdo
	Moderadamente de acuerdo
	Ni de acuerdo ni en desacuerdo
	Moderadamente en desacuerdo
	Totalmente en desacuerdo
Es necesario que los usuarios tengan experiencia previa en el manejo de tableros de análisis:	Totalmente necesario
	Moderadamente necesario
	Ni necesario ni innecesario
	Moderadamente innecesario
	Totalmente innecesario
El análisis y visualización bibliométrica de los grupos de investigación puede ayudar a los actores del Sistema Regional de Ciencia, Tecnología e Innovación del Cauca:	Siempre
	Muchas veces
	Algunas veces
	Pocas veces
	Nunca
¿Qué es lo que más te gustó del *Dashboard*?	Respuesta: _____
¿Qué es lo que menos te gustó del *Dashboard*?	Respuesta: _____
¿Cómo mejorarías el *Dashboard*? Algo que te gustaría que tuviese	Respuesta: _____

Tabla anexos 13. Cuestionario de evaluación. Fuente propia.

Lista de individuos encuestados

- Manuel Fernando Peláez, Investigador *ECoS-CTeI*
- Hendrys Fabian Tobar, Investigador
- Yenny Magaly Castrillón Bolaños, Investigadora y Secretaria *GIMAI*
- Guefry Agredo Mendez, Docente
- Lyda Patricia Mosquera Sánchez, Docente e Investigadora
- Carlos Alberto Rengifo Ruiz, Investigador de Comité de cienciometría
- Carmen Vargas Zuluaga, Apoyo *VRI*
- Adriana Milena Hurtado Montoya, Apoyo *VRI*
- John Yanza, Coordinador de Propiedad Intelectual *VRI*
- Marisol Muñoz Ordóñez, Jefe de División de Gestión de Investigación *VRI*
- Vanessa Ramos, Apoyo a Grupos de Investigación *VRI*
- Elena Rodríguez, Apoyo *DAE* y Coordinadora de Emprendimiento
- Paola Andrea Arciniegas Grijalba, Investigadora
- Gerardo Andrés Torres Rodríguez, Director de *GIMAI*
- Eduardo Rojas Pineda, Codirector *ECoS-CTeI*
- Germán Antonio Arboleda Muñoz, Investigador *ECoS-CTeI*
- Cesar Gómez, Investigador *ECoS-CTeI*
- Jader Marcel Arrechea Castillo, Docente
- Álvaro Rendon Gallón, Investigador *ECoS-CTeI*

Respuestas del cuestionario

En esta sección, se presenta los resultados obtenidos a partir de la encuesta realizada para a evaluación del *Dashboard*. La encuesta fue diseñada con el propósito de recopilar datos relevantes y opiniones de los participantes sobre la utilidad y funcionalidad de la herramienta desarrollada. La hoja de cálculo que alberga los resultados de la encuesta se encuentra disponible en el siguiente enlace:

https://docs.google.com/spreadsheets/d/1eZQeFBz1iyopCk80VbU_ZR4bCwGnh4TLHxOG6HdKo5s/

G. ARTÍCULOS DE INVESTIGACIÓN EN PROCESO DE REVISIÓN PARA LAS REVISTAS “INGENIERÍA E INNOVACIÓN” Y “*PUBLICATIONS*”

Uno de los artículos presenta los resultados correspondientes al Módulo Extractor *CVLAC-GrupLAC* para la presente investigación. Se encuentra en proceso de revisión por la “Revista Ingeniería e Innovación” de la Universidad de Córdoba, Colombia. El otro artículo presenta los Módulos Extractor *Scopus* y *Dashboard* junto con los resultados generales del trabajo de grado, y se encuentra en proceso de revisión por la revista “*Publications*” calificada como A2 en Publindex y con factor de impacto JCR de Q1. Los dos artículos enviados siguen las directrices de las mencionadas revistas y se pueden encontrar en el siguiente enlace:

https://drive.google.com/drive/folders/1HuNLUxk3tsYh2AgedxQ9xS8tKs8DpS-U?usp=sharing

H. ANÁLISIS GENERAL DE GRUPOS DE INVESTIGACIÓN EN EL CAUCA

La presente sección muestra un análisis general de los grupos de investigación en el departamento del Cauca utilizando el sistema desarrollado en el trabajo de grado. Para ello se utilizan datos extraídos por medio de los módulos extractores a fecha de junio de 2023, y se desarrolla un análisis de las principales circunstancias observadas en el panorama de la región con base en la actividad científica de los grupos.

En primera instancia, se toman los 10 grupos con mayor cantidad de registros en todo el departamento y se analiza su actividad con base en todos los tipos de productos manejados en la sección *GrupLAC* del *dashboard*. En la Figura anexos 5 se observa la producción anual de los grupos hasta el año 2023. Se observa muy poca actividad de los grupos antes del año 2000.

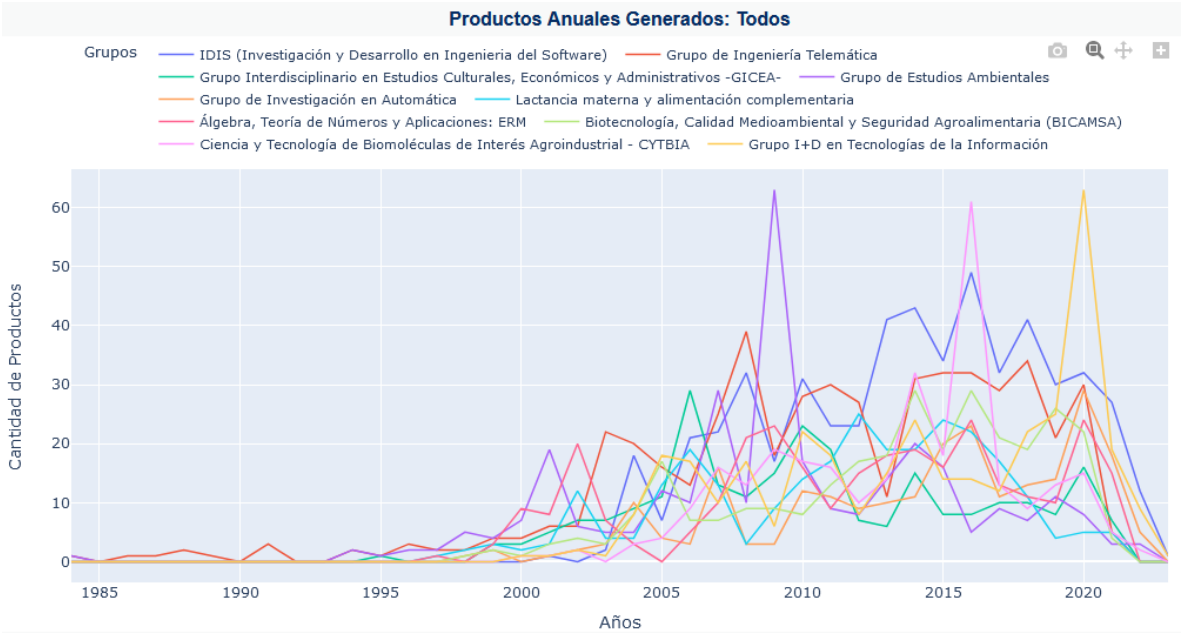


Figura anexos 5. Serie de tiempo comparativo para la producción científica de algunos grupos de investigación filtrados. Fuente propia.

Entre los años 2000 y 2010 se puede notar un aumento generalizado en la productividad de los grupos, llegando algunos de ellos a picos en su historial productivo tales como el Grupo de Estudios Ambientales, el Grupo de Ingeniería

Telemática y el Grupo *GICEA*. Esto se puede apreciar en la Figura anexos 6. La mayoría de los grupos mantuvo menos de 20 productos anuales y la otra parte entre 20 y 65 productos anuales.

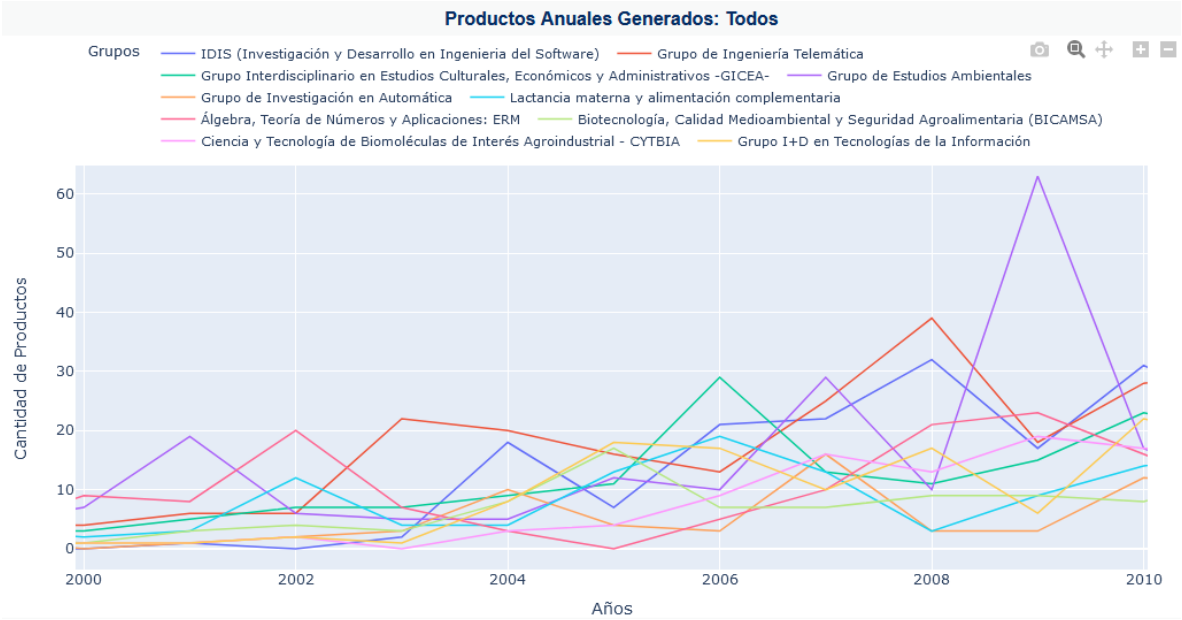


Figura anexos 6. Serie de tiempo comparativo para la producción científica de algunos grupos de investigación filtrados en un rango de fecha. Fuente propia.

Entre los años 2010 a 2023, según la Figura anexos 7, también se observaron picos de productividad para los grupos I+D en Tecnologías de la Información y *CYTBIA*, además se observa que los datos registrados sugieren un descenso importante en la productividad después del 2021. No obstante, este comportamiento puede estar relacionado con el no registro de la información lo cual suele ser estimulado en cada convocatoria para la categorización de grupos de *Minciencias*, siendo la última convocatoria en 2021. Durante el rango de años, la productividad fue más elevada y constante que el rango anterior.

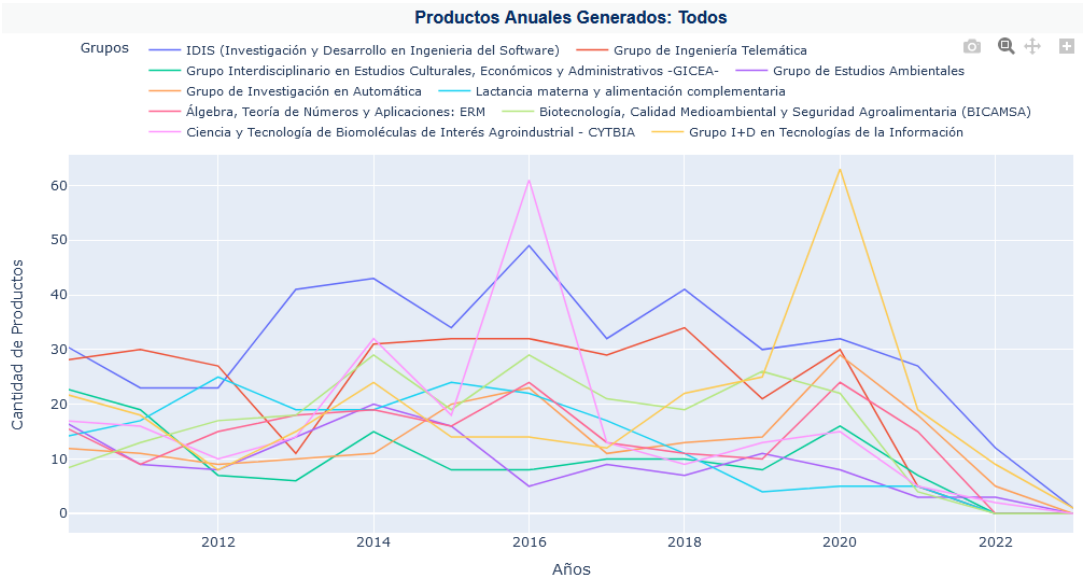


Figura anexos 7. Serie de tiempo comparativo para la producción científica de algunos grupos de investigación filtrados en un rango de fecha. Fuente propia.

En la Figura anexos 8, se observa que el tipo de producto más generado para cada grupo es el artículo de investigación. También se observan proporciones que ocupan parte de la productividad referentes a libros y capítulos de libros. La generación de prototipos, empresas tecnológicas, diseños industriales, plantas piloto en general es baja y de esta pequeña cantidad son responsables los grupos de áreas de ingeniería. Estos grupos también generan la mayor cantidad de software y otros productos tecnológicos. Los cursos y programas de maestría y doctorado son los productos menos generados, sin embargo, cuentan con gran peso y todos los grupos analizados cuentan con contribuciones de este tipo. Esto se puede evidenciar de mejor manera en la Figura anexos 9, donde se observa mayor abundancia en cursos de maestría y cursos de doctorado. La contribución más grande en programas de doctorado ha sido por parte de los grupos *IDIS y ERM*.

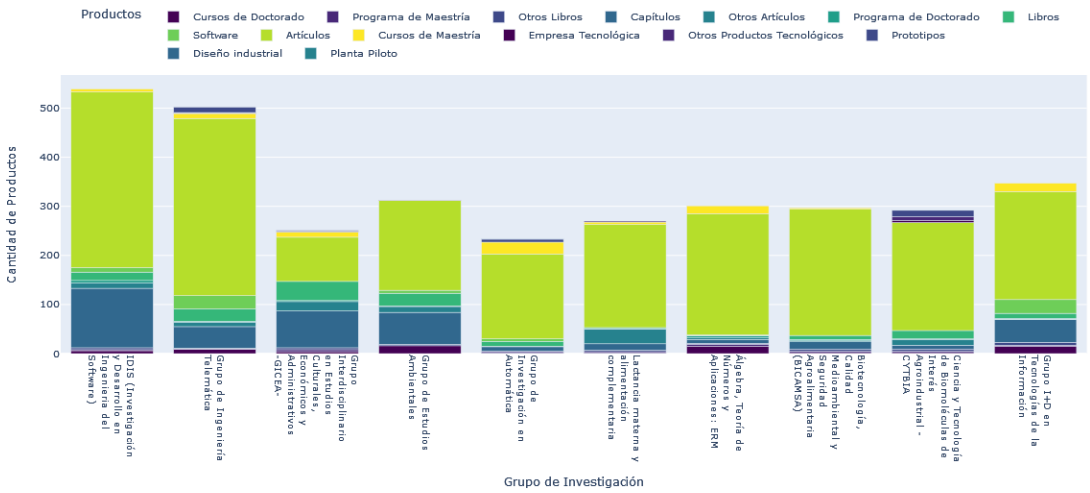


Figura anexos 8. Gráfico de barras comparativo para el tipo de producción científica de algunos grupos de investigación filtrados. Fuente propia.

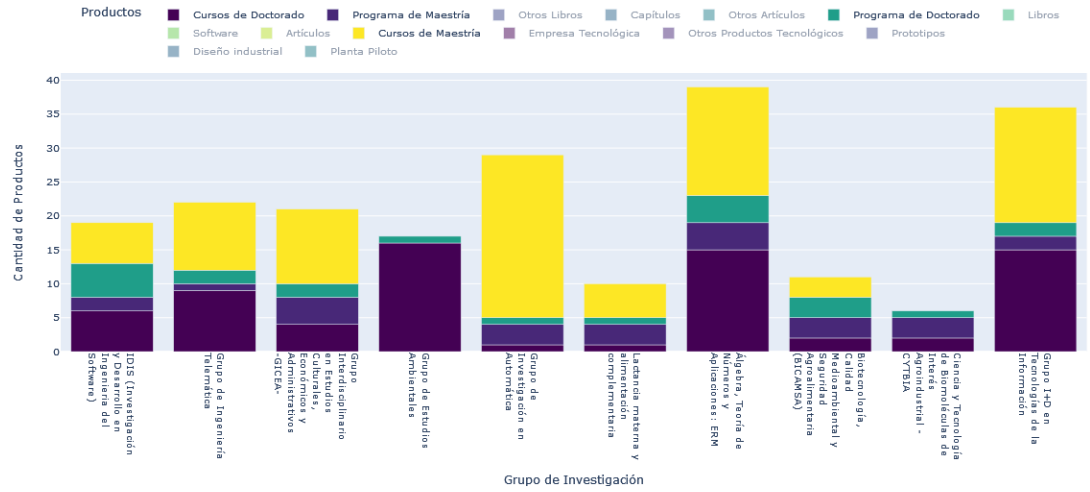


Figura anexos 9. Gráfico de barras comparativo para una selección de tipos de producción científica de algunos grupos de investigación filtrados. Fuente propia.

En las Figuras anexos 10, se observa la comparación de indicadores de consistencia, *PPA, PPUA y PG*. Con base a los registros presentes hasta la fecha, el grupo más consistente y productivo ha sido el grupo *IDIS* por presentar la mayor consistencia, es decir, la menor variación en la publicación de productos durante lapsos anuales dentro de su periodo de actividad, así como su promedio por año y cantidad de productos totales. Adicionalmente, es el grupo que ha contribuido un mayor porcentaje en la generación de productos en todo el departamento en los

últimos 3 años. Otros grupos como I+D en Tecnología e Información y el grupo de investigación en automática, también han tenido una participación destacada. Dado que estos grupos presentan porcentajes abajo del 5% a pesar de ser los más participativos, se tiene que la mayoría de los productos generados en los últimos años han sido aportados por muchos grupos en pequeñas cantidades, teniendo en cuenta que en el Cauca hay un total de 118 grupos.

Figura anexos 10. comparación de indicadores para grupos de investigación filtrados. Consistencia **(a)**, PPA **(b)**, PPUA **(c)** y PG **(d)**. Fuente propia.

Siguiendo con la consistencia, se tiene que una consistencia de 1 se da cuando la desviación estándar de productos y la media aritmética de productos es igual, lo que se considera una dispersión alta de los datos. Un valor menor que 1 para la consistencia, indica una variación mayor que el promedio en los productos anuales. De igual modo, un valor mayor que 1 para la consistencia, indica que la variación es menor que el promedio de los productos anuales, lo cual es deseable. Por lo tanto, la mayoría de los grupos son relativamente consistentes, pero este factor podría mejorar. En cuanto al *PPA*, la mayoría de los grupos están alrededor de los 10 productos anuales. Por último, la cantidad total de productos de estos grupos va desde los 234 hasta los 539, siendo los grupos *IDIS* e Ingeniería telemática los que más han aportado.

Posteriormente, Se realiza el análisis sobre la sección de *Scopus*, tomando los datos emparejados en la base de datos generada para registros de *Scopus* en el Departamento del Cauca. De nuevo se toman los 10 grupos que presentan más registros para todos los tipos de productos de *Scopus*. En general se cuenta con menos registros dado que sólo la producción de más alto impacto ingresa en *Scopus*, además no todos los grupos son visibles. La mayor actividad se tiene entre el periodo de 2015 a 2023 como se observa en la Figura anexos 11. Algunos de los grupos destacados en la sección de *GrupLAC* también se observan aquí, siendo los grupos de ingeniería telemática, *IDIS* e *I+D*, los que presentan los picos más altos de productividad en los primeros meses de los años. La mayor cantidad de publicaciones por parte de todos los grupos se da en el mes de enero, y algunas publicaciones se esparcen entre los otros meses. También se observa que la mayoría de los grupos publican alrededor de 10 productos anuales y sólo unos cuantos más de esta cantidad.

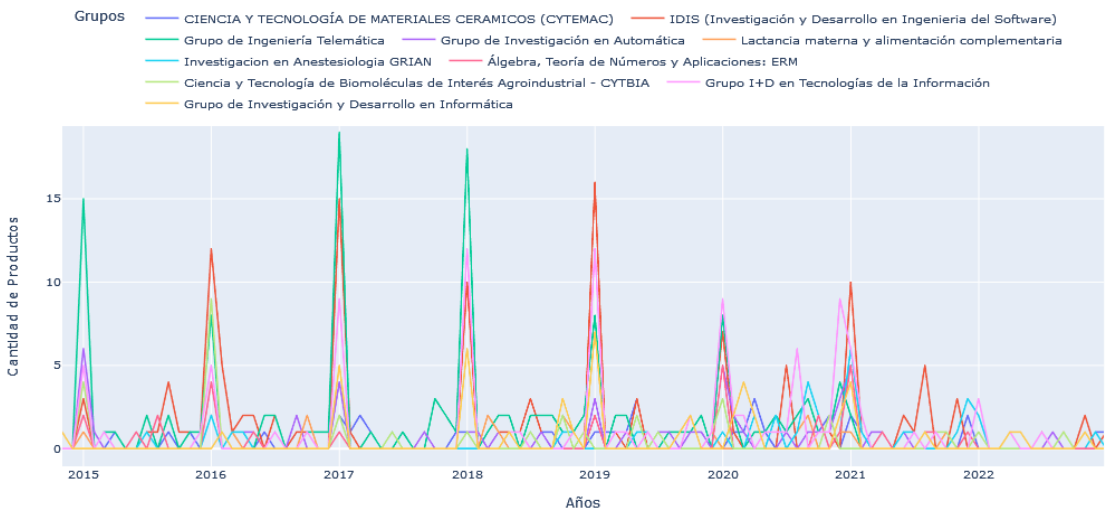


Figura anexos 11. Serie de tiempo comparativo para la producción científica de algunos grupos de investigación filtrados y emparejados con *Scopus*. Fuente propia.

Similar a la sección de *GrupLAC*, se nota un gran descenso en la producción de los últimos años por parte de todos los grupos con altibajos que terminan en 2022 y 2023, lo cual también puede estar relacionado con el motivo expuesto sobre las convocatorias de *Minciencias*. Por otro lado, en la Figura anexos 12 se observan los principales productos generados teniendo a los artículos de investigación y los documentos de conferencia en su mayoría. Se tienen muy pequeñas cantidades para otros tipos de productos y no se observan otros como libros.

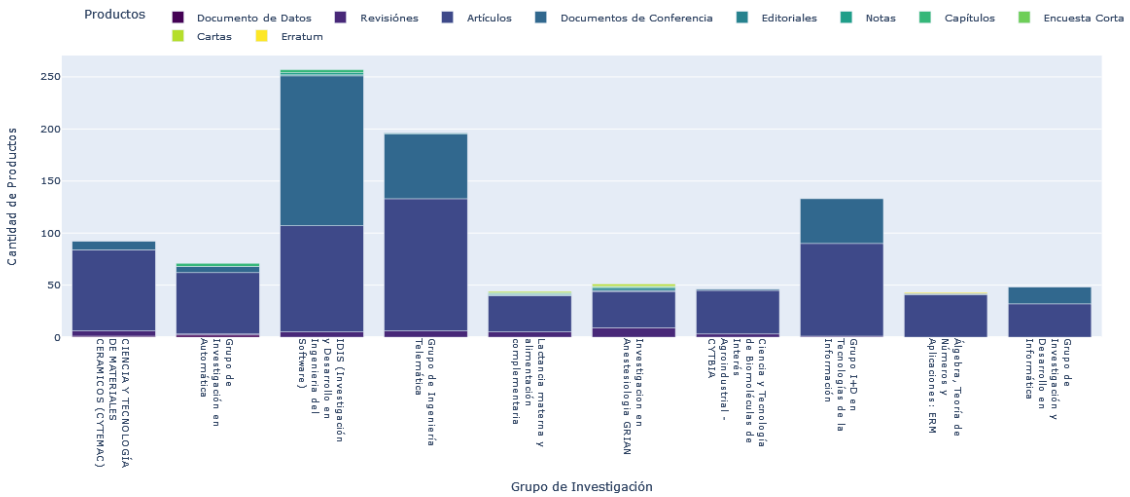


Figura anexos 12. Gráfico de barras comparativo para el tipo de producción científica de algunos grupos de investigación filtrados y emparejados con *Scopus*. Fuente propia.

En la Figura anexos 13 se contrastan los indicadores de calidad e impacto más estrictos para los grupos. Los grupos de ingeniería telemática y de lactancia materna para y alimentación complementaria, presentan los mejores promedios de citaciones por producto. Para el índice h1, los valores más grandes los presentan los grupos de telemática, *IDIS y CYTEMAC*, destacando su productividad e impacto a la vez. En cuanto al índice h2 que tiene en cuenta el desempeño de cada integrante, los grupos de telemática e I+D cumplen con los valores más altos. La mayoría de los grupos tienen un promedio de 5 citaciones por producto y la parte más pequeña presenta un valor superior a 10. En cuanto al índice h1, la mayoría presentan valores alrededor de 10. Por último, la mayoría de los grupos presenta un h2 debajo de 4 y un par se destaca de los demás.

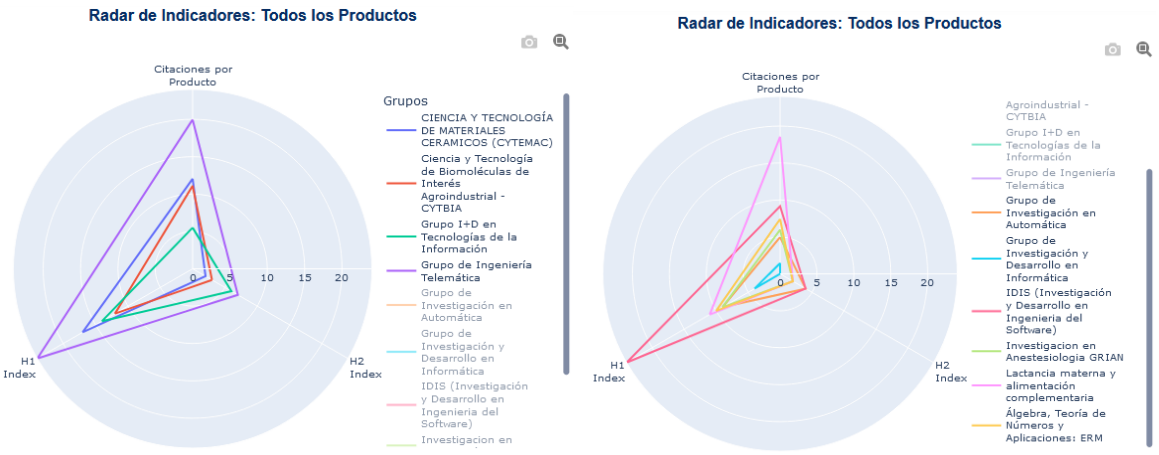


Figura anexos 13. Gráficos de radar comparativo para la producción científica de algunos grupos de investigación filtrados y emparejados con *Scopus*. Fuente propia.

El mapa de calor según las citaciones en la Figura anexos 14 permite conocer a mayor detalle las cantidades, concentrándose la mayoría de ellas en artículos y documentos de conferencia. Se puede notar que el grupo de telemática tiene mayor cantidad total de citaciones en artículos, seguido por *IDIS, CYTEMAC* y Lactancia materna. El grupo *IDIS* presenta más citaciones que los otros grupos en documentos de conferencia. También existen cantidades de citaciones considerables para revisiones por parte de algunos grupos y unas pocas en otros tipos de producto como capítulos o documentos de datos.

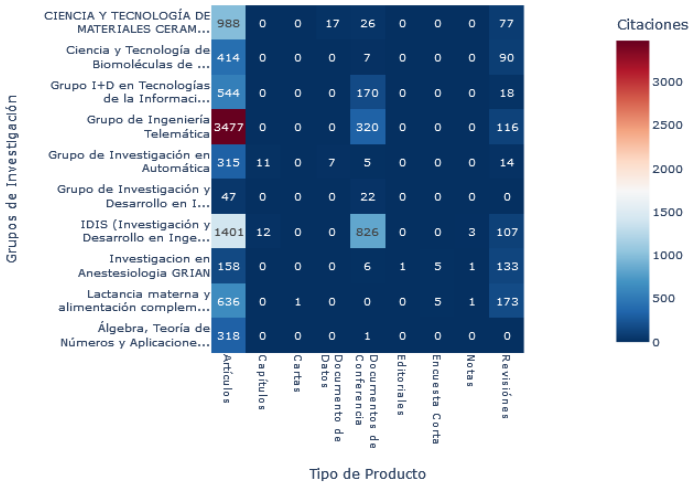


Figura anexos 14. Mapa de calor comparativo para la producción científica de algunos grupos de investigación filtrados y emparejados con *Scopus*. Fuente propia.

A partir de la distribución en la cantidad de citaciones que se genera de cada producto para los grupos, como se observa en la Figura anexos 15, se nota que casi todas las citaciones de los productos están por debajo de 50 para todos los grupos de investigación considerados. Esto se debe a la presencia de *outliers* o valores atípicos que exceden los valores cotidianos, teniendo el producto más citado un valor de 578 para el grupo de ingeniería telemática.

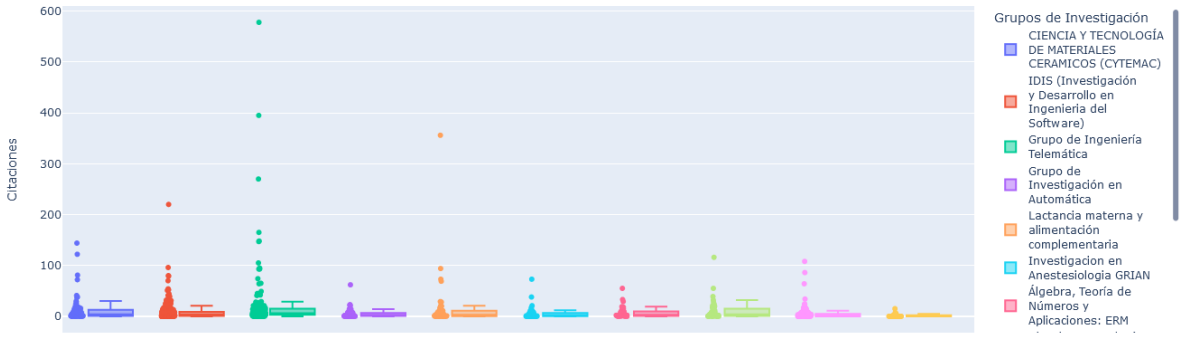


Figura anexos 15. Gráfico de cajas comparativo para la producción científica de algunos grupos de investigación filtrados y emparejados con *Scopus*. Fuente propia.

En la Figura anexos 16, se observan más claramente las cantidades de citaciones para la mayoría de los productos. Se observan cuartiles y medias variadas entre los diferentes grupos, notando valores alrededor de 10 citaciones en su mayoría, por lo

que la mayoría de los productos generados por los grupos tienen pocas citaciones en revistas de alto impacto o no tienen citaciones. Sin embargo, es de destacar que otra parte de los productos logra rebasar los valores esperados y recoger más de 15 citaciones, o hasta cientos de ellas.

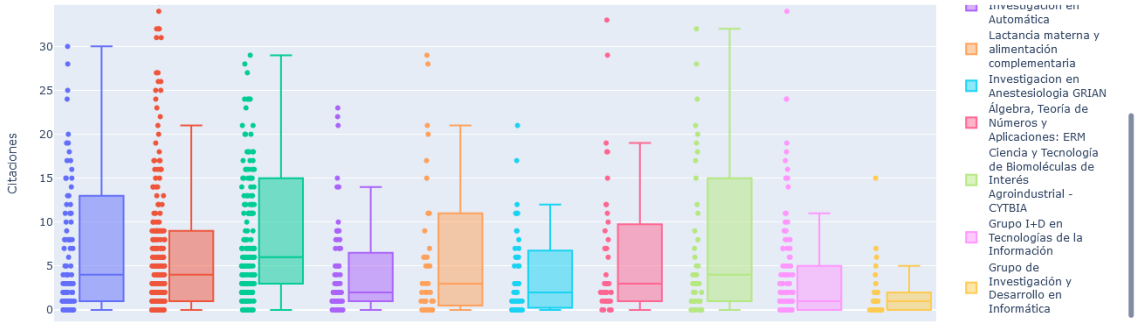


Figura anexos 16. Gráfico de cajas escalado, comparativo para la producción científica de algunos grupos de investigación filtrados y emparejados con *Scopus*. Fuente propia.

Una red de colaboración para el grupo *IDIS*, que fue uno de los más destacados, se observa en la Figura anexos 17, notando su colaboración más grande con el grupo de telemática. No obstante, se trata de una cifra pequeña considerando la productividad reconocida para ambos grupos, aunque cabe señalar que el emparejamiento de los datos con sus grupos también depende de la información registrada en *GrupLAC* por lo que esta cifra puede ser mayor.

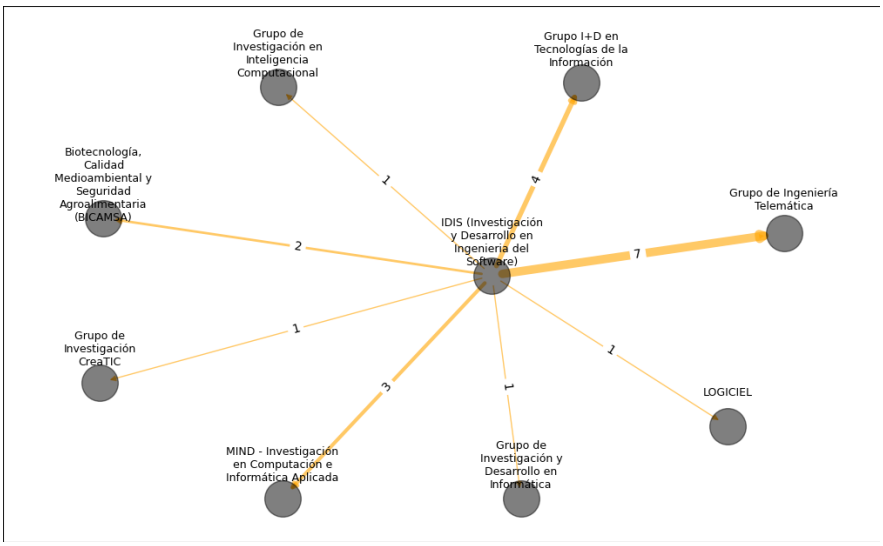


Figura anexos 17. Red de colaboración comparativo entre algunos grupos de investigación filtrados. Fuente propia.

A partir del análisis hecho entre ambas fuentes de datos se puede concluir que existen patrones en la actividad de los grupos a lo largo de los años que indican un incremento considerable de la productividad en los años iniciales de este siglo, con una estabilización que finalmente presenta un importante descenso en la actividad, sin embargo, se señala que es posible que esto se deba a los estímulos que se generan a partir de los procesos de convocatoria y categorización de grupos. También se notó que algunos grupos muestran una mayor presencia en *GrupLAC* sobre su producción en general, pero al observar su actividad en Scopus se evidencia un contraste en cuánto al impacto y productividad en bibliografía de alto impacto por parte de los grupos, resaltando la presencia de algunos de estos grupos en la base de datos de *Scopus* que no se pudo notar a través de *GrupLAC*. Se midió el desempeño de los grupos con base en los indicadores bibliométricos señalando algunos aspectos de productividad, impacto, participación y actividad en un modo comparativo. Finalmente, se conocieron aspectos de la distribución de citaciones presentes en los productos generados por los grupos, analizando aspectos de productividad e impacto señalando donde se concentran la mayoría de los datos y que valores máximos han alcanzado. Cabe mencionar que este análisis se puede extender para explorar otras condiciones de los datos y extraer conocimiento de otros elementos y grupos sobre enfoques más específicos disponibles en el *dashboard*. También es posible extender enfoques de mayor magnitud en versiones futuras de la herramienta según intereses que se identifiquen a partir de los roles o material disponible.