Trabajo Fin de Máster

Análisis de datos bibliométricos

Klever Azael Rumipamba Chicaiza

Máster en Técnicas Estadísticas

Curso 2024-2025

Propuesta de Trabajo Fin de Máster

Título en galego: Análise de datos bibliométricos
Título en español: Análisis de datos bibliométricos
English title: Bibliometric data analysis
Modalidad: Modalidad A
Autor: Klever Azael Rumipamba Chicaiza, Universidade da Coruña
Directores: Rubén Fernández Casal, Universidade da Coruña; Javier Tarrío Saavedra, Universidade da Coruña
Breve resumen del trabajo: La bibliometría es una herramienta clave para analizar la producción, evolución y estado de la actividad científica. Para ello, se requieren datos bibliométricos, cuya obtención y tratamiento representan retos, especialmente por las barreras de acceso y compatibilidad. Estos datos pueden provenir de fuentes abiertas o comerciales, pero su volumen y complejidad hacen necesario el uso de herramientas informáticas. Sin embargo, dichas herramientas no siempre garantizan formatos homogéneos o resultados comparables. Por tanto, el objetivo de este trabajo es revisar y comparar diferentes opciones para la obtención y análisis de datos bibliométricos mediante herramientas informáticas que implementan diversas técnicas estadísticas, aplicándolas al Sistema Universitario de Galicia (SUG).
Otras observaciones:

Don Rubén Fernández Casal, Profesor de la Universidade da Coruña, y don Javier Tarrío Saavedra, Profesor de la Universidade da Coruña, informan que el Trabajo Fin de Máster titulado

Análisis de datos bibliométricos

fue realizado bajo su dirección por don Klever Azael Rumipamba Chicaiza para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal. Además, Don Rubén Fernández Casal, don Javier Tarrío Saavedra y don Klever Azael Rumipamba Chicaiza

√ sí ☐ no

autorizan a la publicación de la memoria en el repositorio de acceso público asociado al Máster en Técnicas Estadísticas.

En A Coruña, a 22 de julio de 2025.

El director:
Don Rubén Fernández Casal

Firmado digitalmente por FERNANDEZ CASAL RUBEN - 33317724Q
Fecha: 2025.07.22 19:24:19 +02'00'

El director:
Don Javier Tarrío Saavedra

TARRIO SAAVEDRA JAVIER - DNI 53164914T
Firmado digitalmente por TARRIO SAAVEDRA JAVIER - DNI 53164914T
Fecha: 2025.07.22 22:00:09 +02'00'

El autor:
Don Klever Azael Rumipamba Chicaiza

Firmado por

Declaración responsable. Para dar cumplimiento a la Ley 3/2022, de 24 de febrero, de convivencia universitaria, referente al plagio en el Trabajo Fin de Máster (Artículo 11, Disposición 2978 del BOE núm. 48 de 2022), **el/la autor/a declara** que el Trabajo Fin de Máster presentado es un documento original en el que se han tenido en cuenta las siguientes consideraciones relativas al uso de material de apoyo desarrollado por otros/as autores/as:

- Todas las fuentes usadas para la elaboración de este trabajo han sido citadas convenientemente (libros, artículos, apuntes de profesorado, páginas web, programas,...)
- Cualquier contenido copiado o traducido textualmente se ha puesto entre comillas, citando su procedencia.
- Se ha hecho constar explícitamente cuando un capítulo, sección, demostración,... sea una adaptación casi literal de alguna fuente existente.

Y, acepta que, si se demostrara lo contrario, se le apliquen las medidas disciplinarias que correspondan.

Agradecimientos

Un saludo a todos aquellos que me conocen, quienes me han ayudado con motivación, conocimiento, experiencia, entre otros. He aprendido, cambiado, y/o reforzado algo de mi ser, por esos y muchos otros motivos, agradezco a todos. Muchas gracias, Dioslepague, Pay, Grazie, Spaciba, ...

Que les vaya bien, Éxitos ...

Índice general

Resumen **xi**

1. Estado del Arte **1**
1.1. Motivación y contexto . . . 1
1.2. Objetivos . . . 1
1.3. Estructura de la memoria . . . 2
1.4. Introducción a la bibliometría . . . 2
1.5. Fuentes de Datos de Publicaciones Científicas . . . 3
1.5.1. Scopus . . . 3
1.5.2. Web of Science (WoS) . . . 4
1.5.3. Google Scholar . . . 5
1.5.4. OpenAlex . . . 6
1.5.5. Crossref . . . 7
1.6. Métodos estadísticos en bibliometría . . . 8
1.6.1. Métricas básicas en bibliometría . . . 9
1.6.2. Coautoría . . . 11
1.6.3. Co-citación . . . 12
1.6.4. Análisis de datos en red . . . 12
1.6.5. Análisis de correspondencias . . . 13
1.7. Herramientas y paquetes en R . . . 17
1.7.1. OpenalexR . . . 17
1.7.2. rcrossref . . . 18
1.7.3. Bibliometrix . . . 19
1.7.4. Scimetr . . . 20

2. Diseño e implementación **23**
2.1. Descripción del caso de estudio . . . 23
2.1.1. Sistema Universitario de Galicia . . . 23
2.1.2. Metodología aplicada al caso de estudio . . . 23
2.2. Obtención y preparación de los datos . . . 23
2.2.1. Obtención de datos desde OpenAlex . . . 24
2.2.2. Obtención de datos desde Crossref . . . 25
2.2.3. Obtención de datos desde WoS . . . 27
2.3. Procesamiento de datos descargados . . . 28
2.3.1. Etiquetado para openalexR y rcrossref . . . 28
2.4. Variables del análisis de correspondencias . . . 28

3. Resultados: caso de estudio del SUG **33**
3.1. Productividad científica de las universidades del SUG. . . . 33
3.1.1. Número de publicaciones por universidad y año. . . . 33

3.2. Impacto de las publicaciones 36
3.2.1. Número total de citas por universidad y por año. 36
3.2.2. Índices bibliométricos fuente-universidad 40
3.3. Patrones de colaboración institucional y de autoría 43
3.4. Áreas temáticas de relevancia y evolución de contenido 50
3.5. Producción del tipo de documento 53

4. Conclusiones y líneas futuras de investigación **59**

Bibliografía **61**

Siglas y Glosario **63**

Apéndice **65**

Resumen

Resumen

Contar con datos representativos en un ámbito de estudio es fundamental en toda labor de investigación. Sin embargo, existen ciertos casos y ámbitos de estudio en los que los datos suelen tener acceso restringido por causas como el costo, la privacidad o la disponibilidad. En el ámbito bibliométrico, la mayoría de las bases de datos son privadas (los datos son un producto comercializado por empresas como Elsevier, Springer, entre otras), lo que representa una importante barrera para la realización tanto de estudios de investigación en este ámbito como de evaluación del desempeño investigador de instituciones e investigadores. En este contexto, los repositorios abiertos como OpenAlex y Crossref se presentan como alternativas a aquellas privadas (y más populares) como *Web of Science* (WoS), *Scopus*, etc. A este respecto, el objetivo del presente trabajo es presentar y realizar una revisión crítica y comparativa de herramientas libres (en términos de acceso a datos y técnicas estadísticas) para el análisis de la cantidad, calidad y visibilidad de la producción científica, tomando como caso de estudio al Sistema Universitario de Galicia (SUG). Con tal fin, se introducen conceptos básicos de bibliometría y fuentes de datos, y se aplican métricas e indicadores usando herramientas como `OpenAlexR` y `Bibliometrix`. Por otro lado, también se han aplicado técnicas de análisis de redes científicas, que proporcionan mapas útiles para caracterizar las áreas de investigación y relaciones de colaboración en el SUG.

Los repositorios abiertos muestran un gran potencial, especialmente OpenAlex, herramienta que ofrece una mayor flexibilidad, volumen de resultados y mejor manejo de identificadores o filtros (el acceso a WoS o Scopus es prohibitivo para muchas universidades debido a su alto coste). En cambio, Crossref presentó limitaciones en la consulta de datos, sobre todo si se compara con otras herramientas, tanto libres como privadas (WoS). De hecho, aunque los repositorios abiertos aún enfrentan desafíos como la estandarización de campos o el acceso a variables clave, su crecimiento y adopción los posicionan como recursos útiles y complementarios frente a las bases comerciales tradicionales (además de integradoras de toda la comunidad científica, independientemente del nivel adquisitivo de las instituciones).

Abstract

Having representative data in a field of study is essential in any research endeavor. However, there are certain cases and areas of study in which data often have restricted access due to reasons such as cost, privacy, or availability. In the field of bibliometrics, most databases are private (data are commercial products marketed by companies such as Elsevier, Springer, among others), which represents a significant barrier to both conducting research in this field and evaluating the research performance of institutions and researchers. In this context, open repositories such as OpenAlex and Crossref emerge as alternatives to the more popular private ones like *Web of Science* (WoS), *Scopus*, etc. In this regard, the aim of this work is to present and conduct a critical and comparative review of open tools (in terms of data access and statistical techniques) for the analysis of the quantity, quality, and visibility of scientific production, taking the Galician University System (SUG) as a case study. To this end, basic concepts of bibliometrics and data sources are introduced, and metrics and indicators are applied using tools such

as `OpenAlexR` and `Bibliometrix`. In addition, scientific network analysis techniques have been applied, providing useful maps to characterize research areas and collaboration relationships within the SUG.

Open repositories show great potential, especially OpenAlex, which offers greater flexibility, volume of results, and better handling of identifiers or filters (access to WoS or Scopus is prohibitive for many universities due to their high cost). In contrast, Crossref showed limitations in data querying, particularly when compared to other tools, both open and commercial (WoS). In fact, although open repositories still face challenges such as standardization of fields or access to key variables, their growth and adoption position them as useful and complementary resources to traditional commercial databases (as well as inclusive of the entire scientific community, regardless of institutions' purchasing power).

Resumo

Dispoñer de datos representativos nun ámbito de estudo é fundamental en calquera tarefa investigadora. Porén, existen certos casos e ámbitos nos que o acceso aos datos adoita estar restrinxido por motivos como o custo, a privacidade ou a dispoñibilidade. No eido da bibliometría, a maioría das bases de datos son privadas (os datos son un produto comercializado por empresas como Elsevier, Springer, entre outras), o que supón unha barreira importante tanto para a realización de estudos neste ámbito como para a avaliación do rendemento investigador de institucións e investigadores. Neste contexto, os repositorios abertos como OpenAlex e Crossref preséntanse como alternativas ás bases privadas (e máis populares) como *Web of Science* (WoS), *Scopus*, etc. Neste senso, o obxectivo deste traballo é presentar e realizar unha revisión crítica e comparativa de ferramentas libres (en termos de acceso aos datos e técnicas estatísticas) para a análise da cantidade, calidade e visibilidade da produción científica, tomando como caso de estudo o Sistema Universitario de Galicia (SUG). Para tal fin, introdúcense conceptos básicos de bibliometría e fontes de datos, e aplícanse métricas e indicadores mediante ferramentas como `OpenAlexR` e `Bibliometrix`. Ademais, tamén se aplicaron técnicas de análise de redes científicas, que proporcionan mapas útiles para caracterizar as áreas de investigación e as relacións de colaboración no SUG.

Os repositorios abertos amosan un gran potencial, especialmente OpenAlex, ferramenta que ofrece maior flexibilidade, volume de resultados e mellor manexo de identificadores ou filtros (o acceso a WoS ou Scopus é prohibitivo para moitas universidades debido ao seu elevado custo). En cambio, Crossref presentou limitacións na consulta de datos, sobre todo en comparación con outras ferramentas, tanto libres como comerciais (WoS). De feito, aínda que os repositorios abertos aínda afrontan desafíos como a estandarización de campos ou o acceso a variables clave, o seu crecemento e adopción colócanos como recursos útiles e complementarios fronte ás bases comerciais tradicionais (así como integradores de toda a comunidade científica, independentemente do nivel adquisitivo das institucións).

Capítulo 1

Estado del Arte

1.1. Motivación y contexto

En Glänzel (2003), se dice que la bibliometría es un campo que estudia la producción científica mediante el análisis de publicaciones académicas. Aunque puede involucrar diversas técnicas estadísticas, en esencia se centra en comprender cómo se produce, se comparte, y se estructura el conocimiento dentro de la comunidad académica. Esta disciplina se ha convertido en una herramienta esencial para evaluar el impacto de la investigación, detectar tendencias emergentes, colaboraciones, entre otros.

Waltman (2016) menciona que a pesar de su utilidad, la bibliometría enfrenta actualmente limitaciones importantes, especialmente en el acceso a datos completos, actualizados y abiertos. Muchas fuentes de información bibliográfica son de acceso restringido (WoS y Scopus, entre otras, que conllevan un alto coste de acceso). Son comerciales o presentan barreras técnicas para su uso masivo (uso de API), lo que dificulta estudios transparentes, reproducibles y accesibles. Se necesita de una suscripción de pago, lo que representa una barrera para el público en general, además de un motivo para acrecentar la desigualdad entre la investigación norte y sur (Malvić et al. 2022). Dado ese motivo, han surgido plataformas de acceso libre, como OpenAlex o Crossref, que representan una alternativa prometedora accesible para todos, independientemente del nivel adquisitivo, para el acceso a datos bibliográficos. Además, el desarrollo reciente de herramientas computacionales de análisis bibliométrico compatibles con el uso de datos bibliográficos a partir de OpenAlex y Crossref, en combinación con paquetes como `Bibliometrix` y `Scimetr` (librerías en `R`), ayuda a hacer cada vez más competitivo su uso frente a WoS o Scopus, las referencias comerciales.

En este contexto, resulta especialmente pertinente realizar una comparación del uso de diversas bases de datos bibliométricas, ya sean libres o de pago, a través de su aplicación al caso de estudio del Sistema Universitario de Galicia (SUG), por su proximidad y relevancia. Esta región cuenta con una comunidad académica activa y una producción científica relevante (Touriñan-López 2025). Evaluar su desempeño en términos de colaboración, productividad y áreas temáticas puede aportar información valiosa para la toma de decisiones institucionales y para visibilizar el papel del SUG en el sistema académico (Vence-Deza 2010).

1.2. Objetivos

El objetivo de este Trabajo Fin de Máster es realizar una revisión de las principales herramientas para el análisis de la cantidad, impacto y visibilidad de la producción científica de investigadores, centros y/o universidades. Todos estos conceptos están relacionados con la calidad de la producción científica (cuantas más citas, cuanto más factor de impacto tenga la revista en la que se publican los artículos,

más verosímil será que el documento publicado tenga una mayor calidad), pero para poder evaluarla completamente, es precisa la revisión cualitativa de profesionales del campo (Aksnes et al. 2019).

Por tanto, el enfoque será:

- En primer lugar, revisar las distintas opciones para la obtención de los datos de publicaciones científicas, incluyendo paquetes de `R` que permiten la descarga directa de bases de datos online (como por ejemplo `openalexR` (Aria et al. 2024) o `rcrossref` (Chamberlain et al. 2014).
- En segundo lugar, revisar el estado del arte de los métodos estadísticos empleados para su análisis, incluyendo los paquetes de `R` que las implementan (como por ejemplo `bibliometrix` (Aria y Cuccurullo 2017) o `scimetr` (Fernandez-Casal et al. 2023)). Las distintas técnicas se ilustrarán empleando `R` con datos de producción científica del Sistema Universitario Gallego (SUG).

1.3. Estructura de la memoria

Los contenidos del documento se detallarán a continuación. En el Capítulo 2 (Estado del arte), se describen las fuentes de información para las Bases de Datos (BBDD), junto con la descripción de las librerías utilizadas, las cuales se describirán, atendiendo a sus capacidades y/o limitaciones en el procesamiento para los datos bibliométricos. Se exponen de manera introductoria los conceptos relacionados con la bibliometría, los cuales serán utilizados a lo largo de este trabajo de investigación, y que permitirán tener una idea o acercamiento general a la materia. Seguido a este, el Capítulo 3 (Diseño e implementación) incluye la metodología llevada para el estudio. En él se abordan aspectos como conexión, consultas, descargas, estructura/propiedades de las librerías utilizadas, entre otros.

El Capítulo 4 (Resultados) muestra resultados de las herramientas bibliométricas y de estadística computacional aplicadas al Sistema Universitario de Galicia (SUG), evaluándolo mediante el cálculo de índices, gráficos, entre otros. Finalmente, se incluye un Capítulo 5 (Conclusiones), que reúne las principales conclusiones y contribuciones del presente trabajo, relacionando los resultados con los objetivos del proyecto de investigación, además de apuntar aquellas posibles líneas de investigación (en relación a la metodología, técnicas aplicadas, estrategias) futuras.

1.4. Introducción a la bibliometría

En Ardanuy (2012), se entiende que la bibliometría, es parte de otra ciencia llamada cienciometría, pero que esta se caracteriza por su enfoque práctico, es decir, se orienta al análisis y cálculo de magnitudes que se puedan considerar cuantificables dentro de la producción o actividad científica. Este enfoque ayuda a caracterizar la investigación realizada por un ente investigador, reflejando también la evolución de su producción científica a lo largo del tiempo. Otros estudios, como el de Tomás-Górriz y Tomás-Casterá (2018), destacan que la bibliometría, al adoptar un enfoque orientado a la aplicación práctica, ha contribuido significativamente a fortalecer la gestión editorial en la producción científica. Esta disciplina proporciona datos relevantes sobre líneas de investigación, instituciones y otros actores clave, permitiendo evaluar el rendimiento académico y su repercusión en la comunidad científica y en la sociedad en general. Será necesario introducir brevemente aquellos conceptos que serán de utilidad a lo largo de este proyecto de investigación, se los nombrarán en varias ocasiones por su participación directa o indirecta con el manejo y/o procesamiento de las BBDD.

Se considerará la clasificación general de tablas que utiliza OpenAlex (Aria et al. 2024), especialmente aquellas de acceso abierto como OpenAlex o Crossref, las cuales buscan proporcionar una estandarización que facilite su gestión. Sitios web correspondientes a bases de datos bibliográficas, como *Web of Science* (WoS), Scopus, Google Scholar, entre otros, implementan enfoques similares. En este contexto, una entidad se entiende como una colección de elementos que comparten atributos o características comunes. Una entidad, por tanto, se entiende como una tabla o matriz de información. Las entidades consideradas por OpenAlex son:

- **Works (Obras)**: La entidad Works(Obras); hace referencia a los documentos académicos tales como: libros, tesis, o documentos académicos en general, los cuales, por las características de una obra, pueden ser ubicados por su DOI o OAID.
- **Authors (Autores)**: En este campo, se encuentran aquellos individuos (o individuo); que crean las Obras (entidad anterior), obras en las cuales se estudian Conceptos (entidad), y que pueden o no estar afiliados a Instituciones (entidad). Entre las características que puede tener un autor, el ORCID es uno de los más importantes.
- **Institutions (Instituciones)**: Entidad que referencia a las universidades u organizaciones que están o tienen una afiliación con los autores (entidad). Las instituciones, a su vez, están vinculadas con la entidad Obras a través de la entidad Autores, pudiendo buscar o identificar una institución a partir de su OAID o su ROR ID.
- **Sources (Fuentes)**: Repositorios en los cuales se albergan o se alojan los trabajos periódicos (revistas, colecciones, entre otros), o trabajos puntuales (actas de congreso, especiales, etc.). Se puede identificar la fuente mediante su OAID o su ISSN-L.
- **Concepts (Conceptos)**: Son las categorías, áreas o etiquetas de agrupamiento, obtenidas de los títulos y/o resúmenes de las publicaciones u obras, y que permiten encontrar/clasificar las publicaciones. Se pueden encontrar las publicaciones mediante OAID o Wikidata ID.
- **Publishers (Editoriales)**: Esta entidad se refiere a organizaciones, empresas y/o centros, los cuales distribuyen documentos académicos, estas editoriales publican varias revistas, en las que se añadirá información extra relacionada con la editorial y/o revista, información que en ciertos casos también puede ser útil.
- **Funders (Financiadores)**: Son las empresas públicas y/o privadas que financian una investigación; usualmente tienen un código asociado, por ejemplo, en Crossref, se localizaría mediante el ROR ID.
- **Geo (Geografía)**: Entidad o tabla que permite reunir obras que se encuentren en un área o zona especificada; para esto se enlazaría con otras tablas o entidades como Autores, Financiadores, etc.

Las entidades o tablas descritas anteriormente son generales y mediante sus características, permiten un manipulado e inmersión en la exploración de las BBDD de mejor manera. Esto se explicará posteriormente cuando se analicen los repositorios para datos bibliométricos.

1.5. Fuentes de Datos de Publicaciones Científicas

En esta sección, se describirán las principales BBDD abiertas (OpenAlex, Crossref), Google Académico (Google Scholar), además de BBDD comerciales como: Scopus y WoS, que gestionan la información; y permiten la descarga de los datos de las publicaciones académicas, para analizarlos posteriormente con herramientas específicamente diseñadas para datos bibliométricos.

1.5.1. Scopus

En su página web oficial se expone que es una base de datos de tipo comercial y que su repositorio alberga información de publicaciones científicas como resúmenes, citas, revisiones, revistas especializadas, libros, patentes y conferencias, entre otros (Baas et al. 2020). Para acceder a sus registros de manera masiva, posee una *Application Programming Interfaces* (API). También proporciona una web interactiva que incluye herramientas que ayudan a los investigadores a conocer el impacto de documentos, autores e instituciones (en este caso a través de la herramienta de benchmarking denominada SciVal). Ofrece métricas como el índice h, el puntaje de citas, etc., que son útiles para evaluar el impacto de publicaciones y autores, además de incluir herramientas para realizar un seguimiento, análisis y/o visualización de resultados en las búsquedas (ver Figura 1.1).

Generalidades: Se necesita una suscripción y una cuenta para poder acceder, sin embargo, se puede ingresar a través de una cuenta institucional o educativa, lo cual puede incluir acceso limitado a ciertas funcionalidades, como la visualización de redes o el seguimiento de citas. Aunque posee una base de datos amplia, esta se ha centrado tradicionalmente en áreas como las ciencias naturales, ciencias sociales e ingenierías. Si bien su alcance es global, existe cierto sesgo hacia publicaciones en inglés y hacia editoriales con alto factor de impacto. Además, su carácter comercial impone restricciones de acceso a usuarios sin suscripción. La disponibilidad de campos específicos puede variar según el tipo de contenido o formato exportado (por ejemplo, RIS o BibTeX).

Scopus proporciona el acceso a datos mediante su SCOPUS API, que permite a desarrolladores y/o analistas automatizar consultas, descargar datos estructurados y realizar análisis. Este acceso es valioso para proyectos que requieren una integración de datos con herramientas como `bibliometrix`, `VOSviewer` (Arruda et al. 2022), o entornos de análisis como `R` o `Python`. Cada registro incluye datos estandarizados atendiendo a diferentes campos como título, resumen, palabras clave, afiliaciones institucionales, país de origen, referencias, DOI y citas recibidas, entre otros. Esta estructura (similar a WoS) facilita la construcción de mapas científicos y redes de investigación (coautoría, co-citación, acoplamiento bibliográfico) y el uso de técnicas estadísticas para la exploración de patrones de producción científica, entre otros análisis.

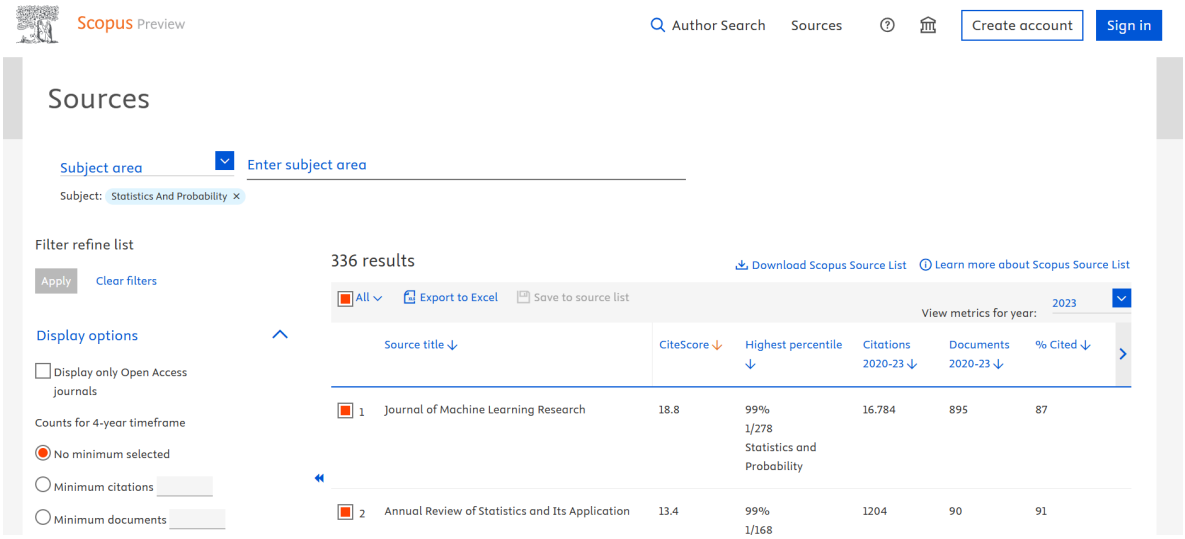

Figura 1.1: Ejemplo búsqueda por área en Scopus

1.5.2. Web of Science (WoS)

Por otra parte, en Pranckutė (2021), se menciona que WoS es una colección o repositorio de datos de varias ciencias (multidisciplinar), siendo una de las más grandes, populares y antiguas. WoS cuenta además con productos como el Journal Citation Reports (JCR), un ranking de revistas realizado, fundamentalmente, en función del valor del indicador denominado factor de impacto, por área de investigación (Krampl 2019). Esta clasificación es una de las más relevantes y comúnmente aceptadas por la comunidad científica para caracterizar la importancia, visibilidad (por extensión, la calidad) de una revista, a través del cuartil del factor de impacto en la que se encuentra. De ahí la importancia de WoS como base de datos de referencia en el ámbito académico.

Generalidades: Se necesita una suscripción y una cuenta para poder acceder. También es posible ingresar mediante una cuenta institucional o educativa. Dispone de una web interactiva para búsquedas y/o consultas automatizadas (véase Figura 1.2), aunque el acceso puede estar limitado según el tipo de

cuenta o institución. Uno de los inconvenientes comunes es la limitación en la descarga de resultados, que suele oscilar entre 500 y 1000 registros por exportación desde la interfaz web. Incluye múltiples colecciones de datos, entre las que destacan el *Social Sciences Citation Index (SSCI)*, *Arts and Humanities Citation Index (A&HCI)*. Cubriendo publicaciones desde 1900 hasta la actualidad. Esto le otorga una profundidad histórica para estudios longitudinales de producción científica. Su cobertura disciplinar es amplia, aunque tradicionalmente más enfocada en publicaciones de alto impacto y en inglés.

Usando la API, es posible hacer búsquedas y descargas de datos en formatos estructurados como TXT, JSON, XML, entre otros. Esto facilita la integración con herramientas de la estadística computacional como `bibliometrix`, etc. Sin embargo, el acceso a esta API depende del tipo de suscripción, y no todas las funcionalidades están habilitadas para usuarios individuales o sin privilegios. Los registros incluyen campos detallados como autores, afiliaciones, resumen, citas, referencias, título de la revista, año de publicación y DOI, entre otros. Esta estandarización favorece los análisis de redes de coautoría, cocitación y acoplamiento bibliográfico. Además, permite el cálculo de indicadores bibliométricos como el índice h o el factor de impacto, entre otras muchas métricas descritas en la Sección 1.6.

A pesar de su prestigio y fiabilidad, WoS parece tener sesgo hacia publicaciones en inglés, así como su limitado acceso abierto. Además, no todas las revistas científicas de regiones emergentes estarían incluidas en su índice principal (Mongeon y Paul-Hus 2016). Su objetivo no sería ser una base de datos exhaustiva como OpenAlex o Google Académico, más bien se orientaría a la selección de fuentes fiables. Estas limitaciones pueden afectar la representatividad de los estudios, pero le darían una ventaja en cuanto a imparcialidad y fiabilidad del contenido, haciéndola ideal para referencias comparativas respecto a otros repositorios.

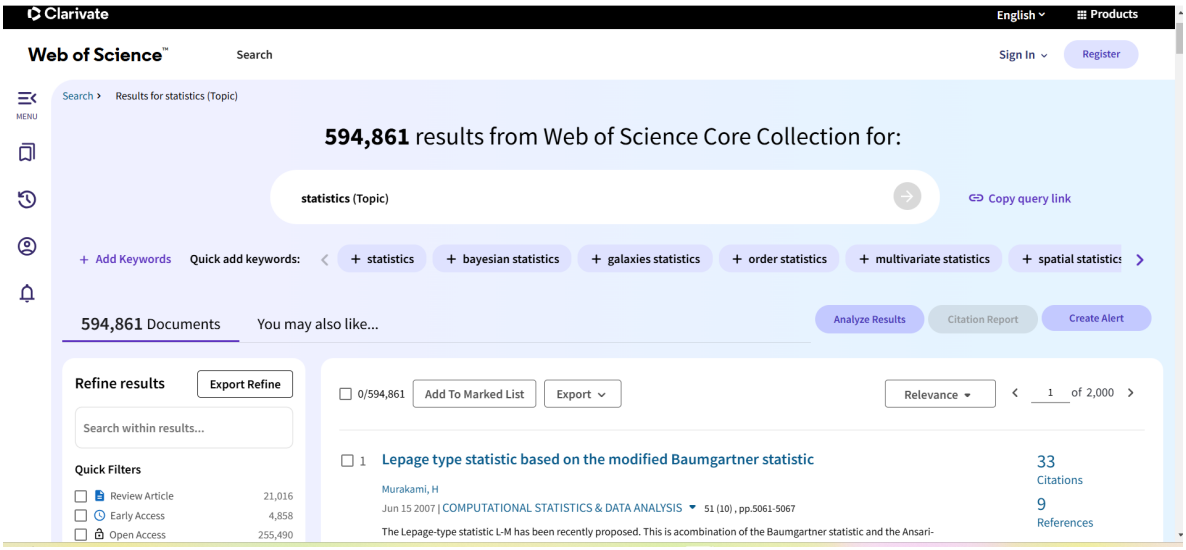

Figura 1.2: Ejemplo búsqueda por tema en WoS

1.5.3. Google Scholar

En Aguillo-Caño (2012), Google Scholar es un instrumento de búsqueda de contenido intelectual (académico en general) que facilita la localización de documentos como artículos científicos, libros, revistas, entre otros. Esta plataforma alberga diversas fuentes académicas, indexando contenidos provenientes de bibliotecas universitarias, editoriales y otras entidades. Entre sus funciones destaca la posibilidad de aplicar filtros como idioma, fecha de publicación o relevancia. En los casos en que el acceso al texto completo está restringido, puede mostrar información parcial como resúmenes, fragmentos, índices o referencias, que permite rastrear o ubicar el documento completo a través de otros medios.

Generalidades: Es totalmente libre o abierta, aunque tiene un motor de búsqueda manual, no funciona como una API, más bien es una web interactiva de búsqueda (Figura 1.3). Proporciona datos como citas, indicadores bibliométricos, perfiles de los autores, artículos relacionados, etc. Existe menor control de la calidad en comparación con bases de tipo comercial o bases especializadas con la gestión de datos académicos. De hecho, puede ocurrir que haya títulos duplicados, autores mal identificados o fechas incorrectas (en una proporción mayor a WoS, por ejemplo), lo cual limita la reproducibilidad de análisis científicos basados en estos datos. A pesar de sus limitaciones estructurales, Google Scholar es ampliamente utilizado cuando los recursos son limitados para acceder a BBDD comerciales como WoS o Scopus. O podría usarse como una primera exploración previa, para tener una idea general de la investigación. Su cobertura es más amplia en términos de idiomas, editoriales y tipos de documentos, lo que la convierte en una herramienta útil.

Debido a que Google Scholar no ofrece una API oficial, su uso en bibliometría automatizada está relativamente restringido. Además, la ausencia de una estructura de datos clara (por ejemplo, identificadores como DOI o ORCID) dificulta la combinación de los resultados con software bibliométrico como `bibliometrix`. Por otro lado, el uso de técnicas de scraping o automatización para recopilar datos parece ser que infringe los términos de uso de Google, lo que genera una barrera técnica y/o legal para su explotación sistemática.

Figura 1.3: Ejemplo búsqueda por tema en *Google Scholar*

1.5.4. OpenAlex

En el Glosario del presente Trabajo Fin de Máster, se proporciona un primer acercamiento general a OpenAlex. En esta seccion, se expondrán más detalles teniendo en cuenta lo expuesto en Borrego-Huerta (2024), además del contenido de su página web oficial. OpenAlex es un repositorio académico abierto, estructurado bajo un modelo de datos basado en 8 entidades (tablas) principales: *Works* (Trabajos), *Authors* (Autores), *Institutions* (Instituciones), *Concepts* (Conceptos), (Waltman 2016), descritos en la Sección 1.4. Esta organización relacional facilita la exploración y análisis de los datos. Cada entidad posee diferentes atributos; por ejemplo, en *Works* (Trabajos), se incluyen campos como título, resumen, autores, afiliaciones, fecha de publicación, citaciones recibidas, referencias, tipo de documento, idioma, DOI (si está registrado en Crossref), etc. Además, la información está enlazada mediante identificadores persistentes (OpenAlex ID, ROR, ORCID, ISSN, etc.). Todo lo anterior convierte a OpenAlex en una alternativa a repositorios de tipo comercial, como lo son Scopus o WoS.

Generalidades: Al no ser comercial, es de tipo abierta, contando además con una web interactiva (Figura 1.4), la cual permite la descarga y/o consultas directas a través de URL estructuradas. Para consultas masivas y análisis más sofisticados, posee una API que admite filtros complejos, ordenamientos, paginación y selección de campos específicos. Por ejemplo, es posible recuperar todos los trabajos de un autor mediante su ORCID, o filtrar trabajos por año, revista, país o número mínimo de citaciones. Además, la API ofrece una estructura estándar en formato JSON, facilitando su exportación y análisis mediante lenguajes como `Python` o `R`. Los límites de paginación son de 200 registros por consulta. El enlazamiento por identificadores primarios o únicos, permite a OpenAlex desenvolverse en el análisis de co-citación, análisis de afiliaciones institucionales, evolución de conceptos, clústeres temáticos, entre otros. Si bien no incluye métricas propias (como CiteScore o factor de impacto), mediante herramientas como `R` y `bibliometrix`, se podrían calcular indicadores como el índice h, impacto medio por autor o por institución, tasas de crecimiento o colaboración internacional, etc. Su sistema de conceptos jerarquizados (similar a Wikidata ID) posibilita análisis temáticos y de evolución disciplinar. OpenAlex representa una alternativa viable y transparente frente a bases comerciales. Pero a diferencia de WoS o Scopus, carece de un sistema de validación formal, lo cual puede implicar ciertas limitaciones para estudios de carácter evaluativo.

Pese a estas desventajas, su adopción está creciendo en la comunidad académica y bibliométrica, hay que tener en cuenta que la expansión o cobertura aún está en desarrollo, por lo que podría ser parcial, especialmente para publicaciones recientes o de editoriales no indexadas por Crossref; de hecho, puede darse que no existan trabajos o publicaciones de alguna entidad, o que no estén incluidos. Aun así, cuenta con una gran cantidad de registros.

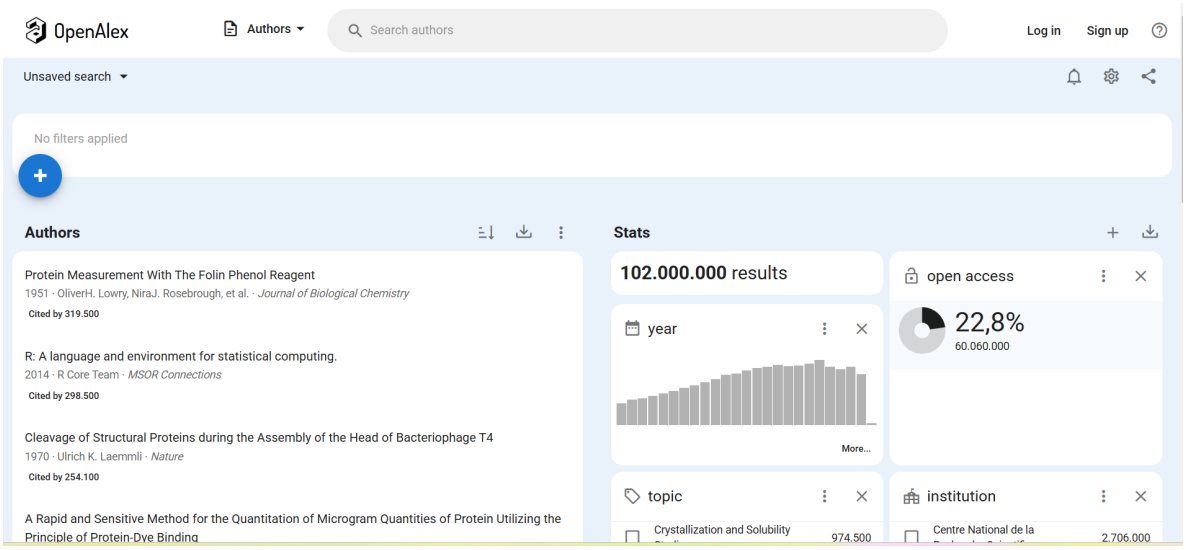

Figura 1.4: Formulario de búsqueda mediante la entidad autores, en la página de OpenAlex (Ejemplo)

1.5.5. Crossref

Tal como se señala en su sitio web oficial, Crossref es una entidad sin fines de lucro cuyo objetivo es facilitar la gestión y conexión del contenido académico (Eck-Nees y Waltman 2022). Para ello, emplea tecnologías como los DOI, que permiten que los materiales científicos como investigaciones, publicaciones u otros trabajos académicos sean fácilmente localizables, citables, accesibles y vinculables en línea. El número de variables que ofrece (título, autor, doi, citas, etc.) es menor, siendo similares a las proporcionadas por OpenAlex.

Generalidades: Los datos disponibles en Crossref son libres; se pueden observar y acceder a ellos

parcialmente mediante su web interactiva (Figura 1.5), mostrando los resultados de aproximadamente las 10 primeras paginaciones. Sin embargo, Crossref no ofrece análisis de citas ni métricas más avanzadas en su formulario de búsqueda. Su función es albergar registros de miles de editores académicos, revistas y actas de conferencias.

Para evitar usar recursos innecesarios en el servidor y poder acceder a los registros completos, se deberá usar la API de Crossref. Aun con el uso de la API, Crossref tiene una cantidad límite de solicitudes que se pueden realizar en un tiempo determinado, siendo actualmente de 50 solicitudes por segundo para aquellos usuarios no registrados, es decir, sin clave de acceso personal. Si se supera el límite, las consultas o solicitudes pueden ser rechazadas por un tiempo. A diferencia de bases de datos como Scopus o WoS, esta plataforma no ofrece métricas propias ni dispone de herramientas integradas para la visualización de datos. Además, en algunos casos, ciertos registros pueden presentar limitaciones, como la ausencia de resúmenes, referencias normalizadas o datos claros sobre afiliaciones institucionales. Por su enfoque principalmente orientado al almacenamiento de información, resulta necesario complementarla con otras fuentes o softwares especializados para llevar a cabo análisis bibliométricos completos.

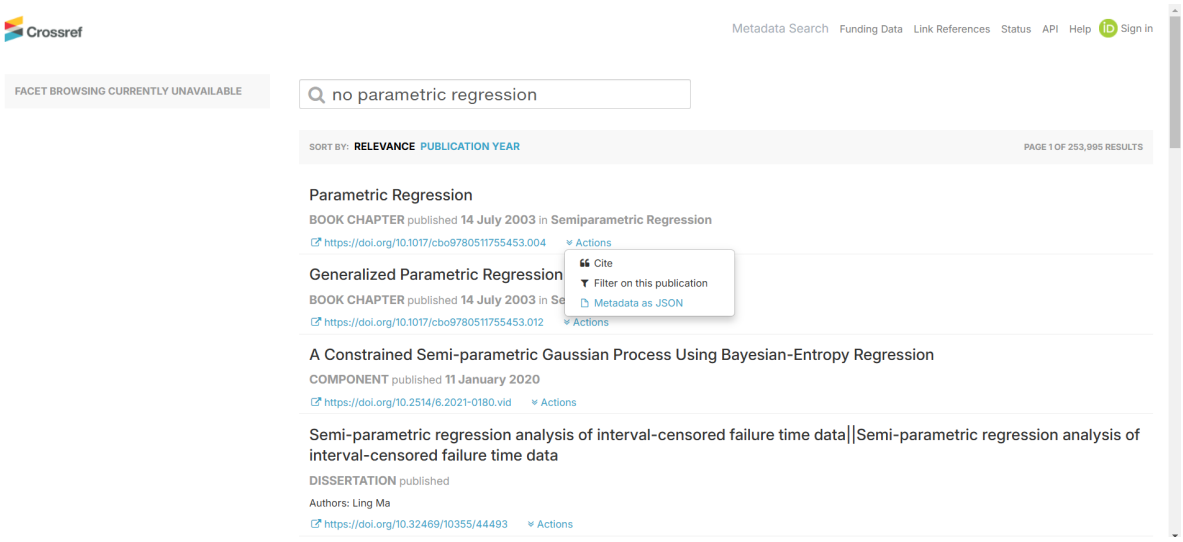

Figura 1.5: Formulario de búsqueda mediante la entidad Works (Obras), en la página de Crossref (Ejemplo)

En esta sección solo se han presentado las características generales de las principales BBDD para el contenido académico, pero existen otras que siguen filosofías similares en la gestión de los datos académicos. En todo caso, en el presente trabajo, se ha optado por describir OpenAlex y Crossref por ser las bases bibliográficas de acceso libre más relevantes en la actualidad.

1.6. Métodos estadísticos en bibliometría

Se menciona en Rubio-Liniers (1999) que los índices o métricas (cuantitativas) utilizados en la bibliometría son herramientas que tienen por objeto analizar y/o medir la producción científica. En esta sección, se darán a conocer los fundamentos o métodos estadísticos que son utilizados dentro de la bibliometría, y que permiten realizar diversos tipos de análisis bibliométricos como mapeos, conteo de citas, productividad e índices de impacto, entre muchos otros. Con el fin de evaluar, medir o caracterizar las publicaciones académicas.

1.6.1. Métricas básicas en bibliometría

Algunas de estas métricas o medidas, presentan sus propias características respecto a otras, pero en conjunto o dependiendo de la investigación, pueden aportar información valiosa. En Hirsch (2005), se mencionan las que se presentarán a continuación.

Número de publicaciones (N_P)

Sea P un conjunto que contiene todos los trabajos pertenecientes a un autor, institución, etc. Considérese i, uno de esos trabajos ($P = \{i : 1 \leq i \leq N_P\}$). La medida o índice N_P, muestra el número de publicaciones que una entidad ha realizado durante un período, para ello se define la siguiente expresión matemática:

$$N_P = card(P) \tag{1.1}$$

Donde:

- $card(P)$: Número total de publicaciones en P (cardinalidad del conjunto P).

Se hace referencia en Gingras (2016), que esta medida es sencilla, pues solo refleja un acercamiento de la producción, pero al medirla en el tiempo, señala una medida de productividad. Por lo tanto, su inconveniente reside en que no evalúa la relevancia o el efecto del contenido académico.

Número de citas (C)

Como su nombre indica, muestra las citaciones recibidas por un trabajo ($i \in P$), dando una idea de la importancia y/o calidad del mismo, es una de las medidas básicas más importantes dentro de la bibliometría, es más en Cooper (2015), se menciona que este índice es uno de los más relevantes dentro de la bibliometría, ya que de este, se basan otros, para encontrar este índice se presenta la siguiente expresión:

$$C = \sum_{i=1}^{N_P} c_i \tag{1.2}$$

Donde :

- c_i: Número de citas recibidas por la publicación i-ésima, en un periodo de tiempo determinado.

Una desventaja podría ser que no siempre están disponibles para algún trabajo, o pueden estar sesgadas por autocitas.

Citas por documento

Este índice mide el promedio de citas recibidas por cada trabajo realizado por una entidad académica (Autor, Institución, País. etc.), se utiliza para evaluar el impacto promedio que tienen las publicaciones en términos de las citas que reciben, su expresión viene dada por:

$$\text{Citas por documento} = \frac{C}{N_p} \tag{1.3}$$

Por su construcción se observa que intenta normalizar las citas por documento, dando una medida relativa de impacto para comparar a las entidades académicas con diferentes niveles de productividad. Sin embargo, puede ser afectado por trabajos con citas altas, haciendo que el promedio se sesgue.

Índice h

Este índice es especial, expresa una medida o cuantificación, en la que combina la productividad y el impacto de un investigador, se dice que un investigador tiene un índice h si ha publicado h artículos con al menos h citas cada uno. Para su cálculo, se ordenan las publicaciones $i \in P$ de forma decreciente según su número de citas:

$$c_{(1)} \geq c_{(2)} \geq \cdots \geq c_{(i)} \geq \cdots \geq c_{(N_P)}$$

Entonces, el índice h se define como el número máximo tal que existen al menos h publicaciones con h o más citas cada una:

$$h = \text{máx}\left\{ i \in P \,\middle|\, c_{(i)} \geq i \right\} \tag{1.4}$$

Donde:

- $c_{(i)}$: Número de citas de la i-ésima publicación ordenada de forma decreciente.

Sin embargo, aunque este índice es ampliamente utilizado, existen críticas hacia él, por su falta de precisión y mal enfoque. Por ejemplo, en Gingras (2016) se menciona que este índice carece de fundamentos matemáticos más rígidos, ya que no posee las características que un buen índice debería tener, un ejemplo de ello es, cuando se menciona que el índice h, se considera incoherente en la manera en que clasifica a investigadores cuyo número total de citas aumenta proporcionalmente, sin que ello se refleje necesariamente en un aumento del índice.

Índice g

Este índice busca complementar al índice h, con base en sus fortalezas y limitaciones. Tiene en cuenta tanto el número de publicaciones como el total de citas acumuladas, otorgando más peso a los trabajos altamente citados. El índice g corrige una de las principales limitaciones del índice h, al considerar no solo cuántas publicaciones son citadas, sino también cuánto han sido citadas en conjunto. Su formulación viene dada bajo el mismo orden decreciente de citas:

Se define como el mayor número tal que las g primeras publicaciones acumulan al menos g^2 citas en total.

$$g = \text{máx}\left\{ i \in P \,\middle|\, \sum_{j=1}^{i} c_{(j)} \geq i^2 \right\} \tag{1.5}$$

Donde:

- $c_{(j)}$: Número de citas de la j-ésima publicación ordenada de forma decreciente.

Entre las ventajas que se pueden mencionar, es que ahora, al considerar el número total de citas, lo hace más sensible a publicaciones altamente citadas, a comparación del índice h. Sin embargo, este índice, por su naturaleza, puede ser menos intuitivo de interpretar que el índice h.

Índice Lotka

Medida bibliométrica que se basa en la ley de Lotka, describe la distribución de la productividad de los autores en un área y/o disciplina académica. En otras palabras, evalúa cuántos autores contribuyen con diferentes cantidades de trabajos en un repositorio. Su formulación es una expresión compuesta:

Ley de Lotka: En términos generales, establece que el número de autores $A(x)$ que publican x trabajos es inversamente proporcional al cuadrado de x, o a una potencia general k, típicamente el cuadrado.

La formulación matemática es la siguiente:

$$\begin{aligned} A(x) &= \frac{A}{x^k} \\ \log(A(x)) &= \log A - k \cdot \log x \end{aligned} \tag{1.6}$$

Donde:

- $N(x)$: Número de autores con exactamente x publicaciones.
- A: Constante de proporcionalidad dependiente del total de autores.
- k: Exponente, comúnmente igual a 2 (ley cuadrática), aunque puede variar según el campo de estudio.

Entre las ventajas y/o desventajas que se pueden mencionar, es que, se puede analizar la concentración de productividad en un campo o área de estudio, identificando autores productivos y/o patrones de contribución. Sin embargo, requiere de un tamaño muestral grande para estimar sus parámetros, los datos duplicados o los errores pueden afectarlo, sesgando así su valor real, se debe considerar que este índice no considera la calidad o impacto de las publicaciones, solo la cantidad.

Aunque dentro del mundo bibliometría existen otras medidas, cada una con sus fortalezas y debilidades, solo se han considerado aquellas medidas básicas con las cuales se pueden construir otros índices más robustos, o con otra finalidad, estos índices básicos están implementados en las librerías (`Bibliometrix` y `Scimetr`), los cuales se van a revisar posteriormente en este trabajo de investigación.

1.6.2. Coautoría

Para poder definir la coautoría, primero se entenderá su palabra raíz, la cual es la autoría. En Drongstrup (2021) se menciona que dentro de la bibliometría, la autoría hace referencia a la conexión, involucramiento o vínculo que tiene el investigador con el trabajo. Por tanto, el término coautoría refiere a cuando dos o más autores colaboran en la elaboración de un trabajo. En bibliometría, el análisis de coautoría permite estudiar las redes de colaboración entre las entidades (autores, instituciones, países, etc.). Se muestra y/o describe los índices o métricas más comunes para analizar la coautoría.

Promedio de autores por artículo

$$IC = \frac{\sum_{i=1}^{N_p} a_i}{N_P} \tag{1.7}$$

Donde:

- a_i: Número de autores en el trabajo i

Esta medida ofrece el promedio de autores por publicación, por lo que es fácil de interpretar y calcular. Permite identificar tendencias en la colaboración según áreas o disciplinas, aunque no representa relaciones directas entre autores.

Índice de colaboración (CI)

Muestra la proporción de trabajos con coautoría respecto al total de los trabajos, se puede ver como:

$$CI = \frac{N_{CA}}{N_P} \tag{1.8}$$

Donde:

- $N_{CA} = \sum_{i=1}^{N_P} \mathbb{I}\{a_i \geq 2\}$: Número de publicaciones con más de un autor (coautoría).

Esta medida, similar a la anterior, permite ver el nivel de colaboración en un conjunto de trabajos, sin embargo, no distingue trabajos con diferente número de coautores.

Grado de coautoría (DC)

Esta medida proporciona el promedio de autores en trabajos con coautoría respecto al total de autores y trabajos, el autor principal no se considera como un coautor en su propio trabajo, es decir que, se enfoca en la colaboración directa de un autor con otros.

$$DC = \frac{\sum_{i=1}^{N_P}(a_i - 1)}{N_P} \tag{1.9}$$

1.6.3. Co-citación

La co-citación es un término utilizado para describir la similitud temática entre dos documentos, artículos o trabajos, en Kleminski et al. (2022), se describen tres tipos básicos:

- **Cita directa:** Si el trabajo B es citado por el trabajo A, se entiende que ambos comparten algo en común, por tanto, ambos están conectados por una red dirigida.
- **Co-citación:** Dos documentos, A y B, están co-citados, si un tercer documento C, cita tanto a A como a B. La frecuencia de co-citación entre A y B es el número de documentos que los citan juntos.
- **Acoplamiento bibliográfico**: Cuando el artículo A es citado por los artículos B, C y D, entonces los conjuntos respectivos: B-C, B-D, C-D, forman relaciones que son parte de la red de acoplamiento bibliográfico no dirigido.

La co-citación se expresaría mediante una matriz M de tamaño $n \times n$, donde n es el número de publicaciones. Cada elemento M_{ij} representa el número de veces que los documentos i y j son co-citados, es decir, que han sido citados conjuntamente por otros documentos.

Frecuencia de Co-citación

La frecuencia de co-citación entre dos documentos i y j se define como el valor M_{ij}, es decir, el número de documentos que los citan juntos.

$$M_{ij} = \sum_{k=1}^{N_P} c_{ki} \cdot c_{kj} \tag{1.10}$$

Donde:

- $c_{ki} = 1$, si el documento k cita al documento i, y 0 en caso contrario.

De manera no directa, en Leung et al. (2017), se expone una representación gráfica que toma como base la matriz anterior, y representa la red de co-citacion.

1.6.4. Análisis de datos en red

El análisis de redes, enfocado a la bibliometría, permite la representación de las estructuras o relaciones entre elementos bibliográficos (universidades, autores, palabras. etc.). En Martínez-Reglero (2018), se da una explicación de los elementos del análisis de redes. Se entiende que los nodos (o vértices) de una red bibliométrica, pueden representar autores, documentos, instituciones, países, entre otros, mientras que las aristas representan relaciones entre ellos, como coautoría o co-citación.

Una red bibliométrica se representaría como un grafo $G = (V, E)$, donde:

- V: Conjunto de nodos, que pueden representar autores, documentos, palabras clave, etc.
- E: Conjunto de aristas o enlaces entre nodos, determinadas según el tipo de relación (coautoría, co-citación, etc.).
- $W : E \to \mathbb{R}^+$: Matriz de pesos, que asigna un valor a cada arista, indicando la intensidad de la relación entre los nodos.

La matriz de adyacencia asociada a la red $W_=(w_{ij})$, representa el peso entre los nodos i y j. En la página oficial de Bibliometrix, se da a entender que esta matriz W, sería una red bipartita, similar a la matriz anterior. A continuación, se define la construcción de las redes, que se usarán con la librería `bibliometrix`:

Red de coautoría Una red de coautoría entre autores, se entiende que es una relación donde los nodos son autores y los enlaces son coautorías. Con base en ello se define:

- Cada nodo $v_i \in V$ representa un autor.
- Existe una arista $(v_i, v_j) \in E$ si los autores i y j han colaborado en al menos una publicación.
- El peso w_{ij} se define como el número de publicaciones que comparten:

$$w_{ij} = \sum_{k=1}^{N_P} \mathbb{I}_k^{(i,j)}, \quad \text{donde } \mathbb{I}_k^{(i,j)} = \begin{cases} 1, & \text{si } i \text{ y } j \text{ son coautores en el trabajo } k, \\ 0, & \text{en otro caso.} \end{cases}$$

Red de co-palabras En esta red, los nodos representarían palabras clave y las aristas indican la co-ocurrencia de dichas palabras dentro de un mismo documento. Esta red permite analizar la estructura temática de un campo de estudio, revelando los conceptos que aparecen frecuentemente juntos. Es decir, identificaría temas o áreas de investigación recurrentes a través de palabras clave o términos en títulos y resúmenes. Permitiendo analizar la evolución temática de un campo.

- Cada nodo $v_i \in V$ representa una palabra clave o término.
- Existe una arista (v_i, v_j) si ambas palabras aparecen juntas en al menos un documento.

$$w_{ij} = \sum_{k=1}^{N_P} t_{ki} \cdot t_{kj}$$

Donde $t_{ki} = 1$ si la palabra i aparece en el documento k, 0 en otro caso.

1.6.5. Análisis de correspondencias

Se menciona en Cuadras-Avellana (2007), que el análisis de correspondencias (AC) se define como una técnica perteneciente ala estadística multivariante, orientada al análisis exploratorio de datos categóricos. Su objetivo es representar gráficamente la asociación entre las categorías de dos variables cualitativas, permitiendo identificar patrones de asociación y diferencias estructurales. Es especialmente útil para analizar tablas de contingencia donde los elementos representan frecuencias de ocurrencia conjunta.

Para empezar a definir el análisis de correspondencias se empezará por introducir parte de sus elementos, características, y limitaciones. Este trabajo de investigación solo considerará el análisis de correspondencias simples, debido a la facilidad de interpretación, construcción, y por estar relacionado con librerías como `bibliometrix()`.

Test χ^2 de independencia

Considérese una tabla de contingencia (tabla de frecuencias), $\mathbf{F} = (f_{ij})_{r\times s}$, con r filas (categorías de X) y s columnas (categorías de Y). Cada elemento n_{rs} representaría la frecuencia observada de la categoría r con la categoría s, y $f_{ij} = \frac{n_{ij}}{n}$ la frecuencia relativa. Ver Tabla 1.1.

Tabla 1.1: Tabla de contingencia con frecuencias relativas $f_{ij} = \frac{n_{ij}}{n}$

$X\backslash Y$	y_1	$\cdots$	y_j	$\cdots$	y_s	$f_{i\bullet} = \frac{n_{i\bullet}}{n}$
x_1	$f_{11} = \frac{n_{11}}{n}$	$\cdots$	$f_{1j} = \frac{n_{1j}}{n}$	$\cdots$	$f_{1s} = \frac{n_{1s}}{n}$	$f_{1\bullet} = \frac{n_{1\bullet}}{n}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_i	$f_{i1} = \frac{n_{i1}}{n}$	$\cdots$	$f_{ij} = \frac{n_{ij}}{n}$	$\cdots$	$f_{is} = \frac{n_{is}}{n}$	$f_{i\bullet} = \frac{n_{i\bullet}}{n}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_r	$f_{r1} = \frac{n_{r1}}{n}$	$\cdots$	$f_{rj} = \frac{n_{rj}}{n}$	$\cdots$	$f_{rs} = \frac{n_{rs}}{n}$	$f_{r\bullet} = \frac{n_{r\bullet}}{n}$
$f_{\bullet j} = \frac{n_{\bullet j}}{n}$	$f_{\bullet 1}$	$\cdots$	$f_{\bullet j}$	$\cdots$	$f_{\bullet s}$	1

Por propiedades de probabilidad y/o conjuntos (aplicados a variables aleatorias), se sabe que la independencia de dos variables aleatorias es equivalente a que la distribución conjunta resulte del producto de las marginales. Esto es importante para poder introducir el concepto de **frecuencias esperadas** dadas por $\frac{n_{i\bullet}n_{\bullet j}}{n}$. Con estos elementos es posible crear el estadístico χ^2:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij} - n_{i\bullet}n_{\bullet j}/n)^2}{n_{i\bullet}n_{\bullet j}/n} \quad \text{(Forma con frecuencias absolutas)}$$

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(f_{ij}n - f_{i\bullet}f_{\bullet j}n)^2}{f_{i\bullet}f_{\bullet j}n} \quad \text{o} \quad \underbrace{\chi^2 = n\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(f_{ij} - f_{i\bullet}f_{\bullet j})^2}{f_{i\bullet}f_{\bullet j}}}_{\text{Forma Simplificada}} \quad \text{(Con frecuencias relativas)} \tag{1.11}$$

Que permitiría contrastar la siguiente prueba de hipótesis:

$$\begin{cases} H_0 : \text{No hay asociación entre } X \text{ e } Y. \ (\textit{Independencia}). \\ \\ H_1 : \text{Hay asociación entre } X \text{ e } Y. (\textit{Dependencia}). \end{cases} \tag{1.12}$$

Bajo H_0 de independencia, el estadístico (1.11) sigue una distribución χ^2 con $(r-1)\times(s-1)$ grados de libertad. Obsérvese que además de permitir contrastar lo expuesto en 1.12, en cierta manera con 1.11, se tendría una medida de asociación entre las categorías y/o niveles de ambas variables, como la V de Cramer o el coeficiente de contingencia. Por tanto, se entendería que el análisis de correspondencias sería un procedimiento el cual busca dar una representación e interpretación de la variabilidad resultante del estadístico χ^2. Dicho esto, si H_0 fuese cierta, no sería razonable llevar a cabo un análisis de correspondencias. En caso contrario, se tendría evidencia estadística de una asociación entre variables, que a través del análisis de correspondencias se estudiará dicha asociación.

Perfiles fila y columna

Para hacer una representación conjunta de la asociación de las variables X e Y. Primero se conseguirán las distancias entre las categorías (filas o columnas). La matriz de frecuencias relativas conjuntas era $\mathbf{F} = (f_{ij})$. Considerando las filas, se trataría de buscar una representación de las r categorías o puntos, en un espacio de menor dimensión, tal que permita apreciar las distancias entre ellas. Si se quisiera medir la distancia entre las filas de $\mathbf{F}$ mediante la distancia euclídea, no sería una buena opción. Considérese el vector que contiene las marginales de X como $m_x = (f_{1\bullet}, \cdots, f_{r\bullet})^{'}$, si ahora se dividen los elementos de la matriz $\mathbf{F}$, por las filas marginales m_x, la matriz resultante $\mathbf{R}$, (matriz de frecuencias relativas condicionadas al total de la fila), matricialmente tendría la forma:

$$\mathbf{R} = \mathbf{D}_r^{-1}\mathbf{F}$$

Con $\mathbf{D}_r$ una matriz diagonal con los elementos de m_x. Cada fila i-ésima de $\mathbf{R}$, es a lo que se conoce como perfil fila i-ésimo (sería también la distribución de Y condicionada a $X = x_i$) Para los perfiles columnas la idea de construcción sería la misma, se utilizará la distribución marginal de Y, con $m_y = (f_{\bullet 1}, \cdots, f_{\bullet s})^{'}$. La Tabla 1.2 contiene los perfiles de fila, junto con la distribución marginal de Y.

Tabla 1.2: Perfiles de fila con frecuencias relativas condicionales $\frac{f_{ij}}{f_i}$

$X \backslash Y$	y_1	$\cdots$	y_j	$\cdots$	y_s	
x_1	$\frac{f_{11}}{f_{1\bullet}}$	$\cdots$	$\frac{f_{1j}}{f_{1\bullet}}$	$\cdots$	$\frac{f_{1s}}{f_{1\bullet}}$	1
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_i	$\frac{f_{i1}}{f_{i\bullet}}$	$\cdots$	$\frac{f_{ij}}{f_{i\bullet}}$	$\cdots$	$\frac{f_{is}}{f_{i\bullet}}$	1
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_r	$\frac{f_{r1}}{f_{r\bullet}}$	$\cdots$	$\frac{f_{rj}}{f_{r\bullet}}$	$\cdots$	$\frac{f_{rs}}{f_{r\bullet}}$	1
Marginal y	$f_{\bullet 1}$	$\cdots$	$f_{\bullet j}$	$\cdots$	$f_{\bullet s}$	1

En teoría, bajo un modelo de independencia, las distribuciones de la variable Y, condicionadas a cada valor x, deberían ser iguales a la distribución marginal de Y. En análisis de correspondencias se compararían todos los perfiles fila con el perfil de fila medio (vector m_y).

Perfiles fila y columna estandarizadas

Obsérvese que con lo desarrollado hasta ahora, cada perfil fila (filas de $\mathbf{R}$), sería un punto en un espacio $\mathbb{R}^s$. Bajo este contexto se buscaría una manera de proyectar dichos puntos en un espacio de dimensión menor, de forma que filas similares estén próximas y las filas diferentes estén alejadas. Se necesitará, por tanto, una distancia y/o medida que permita tener comparaciones razonables teniendo en cuenta la frecuencia relativa de cada nivel de Y. Para ello considérese como distancia entre las filas r_a y r_b de $\mathbf{R}$, lo siguiente:

$$d^2(r_a, r_b) = \sum_{j=1}^{s} \frac{(r_{aj} - r_{bj})^2}{f_{\bullet j}} = \sum_{j=1}^{s} \left(\frac{f_{aj}}{f_{a\bullet}} - \frac{f_{bj}}{f_{b\bullet}} \right)^2 \frac{1}{f_{\bullet j}} = \sum_{j=1}^{s} \left(\frac{f_{aj}}{f_{a\bullet}\sqrt{f_{\bullet j}}} - \frac{f_{bj}}{f_{b\bullet}\sqrt{f_{\bullet j}}} \right)^2$$
$$\underbrace{d^2(r_a, r_b) = (r_a - r_b)' \mathbf{D}_c^{-1} (r_a - r_b)}_{\text{Forma Matricial}} \tag{1.13}$$

La expresión 1.13, se conoce como distancia χ^2. Donde D_c sería una matriz diagonal con los elementos del vector m_y, es decir que permitiría estandarizar las filas de $\mathbf{R}$, por ejemplo una fila estandariza sería:

$$r_i^e = \left(\frac{f_{i1}}{f_{i\bullet}\sqrt{f_{\bullet 1}}}, \ldots, \frac{f_{is}}{f_{i\bullet}\sqrt{f_{\bullet s}}} \right)' \quad i \in \{1, \ldots, r\} \tag{1.14}$$

Lo expuesto en 1.14, se conoce como **perfil fila estandarizado**. Con las formulaciones anteriores obsérvese que 1.13, coincidiría con la distancia euclídea entre los perfiles fila estandarizados correspondientes. Para llevar lo anterior a una forma matricial, considérese una transformación en $\mathbf{R}$:

$$\begin{aligned} \mathbf{X} &= \mathbf{R}\mathbf{D}_c^{-1/2} = \mathbf{D}_r^{-1}\mathbf{F}\mathbf{D}_c^{-1/2} \\ \mathbf{X} &= \left(\frac{f_{ij}}{f_{i\bullet}\sqrt{f_{\bullet j}}} \right)_{i,j} = \begin{pmatrix} r_1^{e'} \\ \vdots \\ r_r^{e'} \end{pmatrix} \end{aligned} \tag{1.15}$$

Análisis de componentes principales para perfiles estandarizados

En la Ecuación 1.15, se tendría una matriz cuyas filas son los perfiles fila estandarizados. Esto con el objetivo de poder proyectarlo hacia un espacio de dimensión menor, de manera que las filas similares aparezcan más próximas en la proyección. Considerando las ponderaciones de las frecuencias marginales de $f_{i\bullet}$, entonces el vector de medias es: $\mathbf{m}_y^e = \left(\sqrt{f_{\bullet 1}}, \ldots, \sqrt{f_{\bullet s}}\right)'$ y la matriz de covarianzas: $\mathbf{\Sigma_r} = \mathbf{X}'\mathbf{D}_r\mathbf{X} - m_y^e(m_y^e)'$, siendo $\mathbf{D_r} = diag(f_{1\bullet}, \cdots, f_{r\bullet})$. Ahora ya se tendrían los componentes necesarios para un análisis de componentes principales, que se traduciría a diagonalizar la matriz de covarianzas $\mathbf{\Sigma_r}$, en este sentido se tendría que $\mathbf{m}_y^e$ es un autovector de $\mathbf{\Sigma_r}$ de autovalor 0, y el resto de autovectores y autovalores de $\mathbf{\Sigma_r}$ y $\mathbf{X}'\mathbf{D}_r\mathbf{X}$ coincidirían.

Volviendo al estadístico χ^2 (1.11), si se lo divide por n, se tendría la **inercia total**:

$$\frac{\chi^2}{n} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(f_{i\bullet}\frac{f_{ij}}{f_{i\bullet}} - f_{i\bullet}f_{\bullet j}\right)^2}{f_{i\bullet}f_{\bullet j}} = \sum_{i=1}^{r} f_{i\bullet} \sum_{j=1}^{s} \left(\frac{f_{ij}}{f_{i\bullet}} - f_{\bullet j}\right)^2 \frac{1}{f_{\bullet j}} = \sum_{i=1}^{r} f_{i\bullet} d^2(r_i, m_y) \tag{1.16}$$

La inercia equivaldría a una medida ponderada de las distancias χ^2 de los perfiles fila a la distribución marginal de Y. La distribución marginal se obtendría como la media de los perfiles fila, por tanto, el estadístico χ^2 sería una medida de variabilidad de los perfiles en torno a su media, en este caso perfiles fila alrededor de m_y. Bajo H_0, el supuesto de independencia (1.12), se esperaría que los perfiles sean similares entre sí y similares a la distribución marginal, dando un valor pequeño de variabilidad, esto medido y/o recogido por el estadístico χ^2.

El análisis de componentes principales de los perfiles estandarizados da lugar a una descomposición de la inercia como la suma de los autovalores:

$$\begin{aligned} \frac{\chi^2}{n} &= \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(f_{ij} - f_{i\bullet}f_{\bullet j})^2}{f_{i\bullet}f_{\bullet j}} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{f_{ij}^2}{f_{i\bullet}f_{\bullet j}} - 1 = \text{traza}\left(X'D_rX\right) - 1 = \text{traza}\left(\Sigma_r\right) \\ \frac{\chi^2}{n} &= \text{traza}(\mathbf{\Sigma}_r) = \lambda_2 + \lambda_3 + \ldots + \lambda_s \end{aligned} \tag{1.17}$$

Por tanto, la proporción de la variabilidad explicada por $(k-1)$ componentes $(k \leq s)$ es $\frac{\lambda_2+\ldots+\lambda_k}{\lambda_2+\ldots+\lambda_s}$

Representación conjunta de los perfiles fila y columna

Para la representación conjunta, se necesitará considerar la siguiente matriz $\boldsymbol{Z} = \left(\frac{f_{ij}}{\sqrt{f_{i\bullet}}\sqrt{f_{\bullet j}}} \right)_{i,j}$, que se relaciona con las expresiones matriciales halladas anteriormente:

$$\mathbf{X} = \mathbf{D_r^{-1}} \mathbf{F} \mathbf{D_s^{-1/2}}, \quad \mathbf{Y} = \mathbf{D_r^{-1/2}} \mathbf{F} \mathbf{D_s^{-1}}, \quad \mathbf{Z} = \mathbf{D_r^{-1/2}} \mathbf{F} \mathbf{D_s^{-1/2}}$$

Por tanto, $\mathbf{X}'\mathbf{D_r}\mathbf{X} = \mathbf{Z}'\mathbf{Z}$, $\mathbf{Y}\mathbf{D_s}\mathbf{Y}' = \mathbf{Z}\mathbf{Z}'$, si $\mathbf{v}$ es un autovector de $\mathbf{Z}'\mathbf{Z}$ con autovalor λ, entonces $\mathbf{w} = \mathbf{Z}\mathbf{v}$ será un autovector de $\mathbf{Z}\mathbf{Z}'$ con autovalor λ. Con esto se vería que el número de autovalores no nulos sería $k_0 \leq min(r, s)$. Las coordenadas para la representación conjunta se obtendrían de la descomposición en valores singulares de $\mathbf{Z} = \mathbf{D_r^{-1/2}} \mathbf{F} \mathbf{D_s^{-1/2}}$

Perfiles fila $= \mathbf{D_r^{-1/2}}\,\mathbf{U_k}\,\mathbf{\Sigma_k}$	Perfiles columna $= \mathbf{D_s^{-1/2}}\,\mathbf{V_k}\,\mathbf{\Sigma_k}$
$\mathbf{U_k}$ autovectores de $\mathbf{Z}\mathbf{Z}'$	$\mathbf{V_k}$ autovectores de $\mathbf{Z}'\mathbf{Z}$
$\mathbf{\Sigma_k} = \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_k})$	(autovalores compartidos)

1.7. Herramientas y paquetes en R

En esta sección se detallarán las herramientas con las que se procesarán los datos bibliométricos de las distintas fuentes de datos abiertos. Para ello se hará uso de software libre como R, debido a sus capacidades como: procesar grandes volúmenes de datos, aplicar métodos estadísticos avanzados, entre otros. Las librerías `openalexR` y/o `rcrossref` permitirán la conexión a los repositorios de las BBDD abiertas. El análisis de datos bibliométricos se describirá con Bibliometrix o Scimetr.

1.7.1. OpenalexR

En Aria et al. (2024), mencionan que openalexR, es una librería que permite interactuar de manera directa con la API de OpenAlex, permitiendo así, consultas y descargas de los datos que se albergan en ese repositorio (publicaciones). Dando la posibilidad de transformar/obtener dichos datos, en un dataframe de `R` (función `oa2df()`), para posteriormente poder procesarlos con la ayuda de librerías como `bibliometrix` u otros. Como se mencionó en la Subsección 1.5.4, OpenAlex, es una alternativa abierta y gratuita a bases de datos comerciales como Scopus o Web of Science, y poder acceder mediante herramientas especializadas como `R`, la hacen una alternativa muy interesante.

Características principales

Entre las diferentes características que ofrece la librería `openalexR`, vease la Tabla 1.3, y las expuestas a continuación:

- Permite el acceso a datos relevantes de publicaciones (autor, resúmenes, títulos, etc.) de OpenAlex,
- Tiene una interfaz amigable con funciones bien documentadas, posee documentación para encontrar la descripción de las distintas funciones que abarca.
- Permite la integración con otros paquetes de `R`.
- Una vez conectadas/descargadas las BBDD, se pueden combinar con herramientas como `bibliometrix` , y se podrían: calcular índices bibliométricos, análisis de co-autoría, co-citación, análisis de co-palabras, clústeres, también se podría un análisis temático y detección de tendencias.

Tabla 1.3: Funciones principales del paquete `openalexR`

Enfoque	Función	Descripción
Consulta de datos	*oa_query()*	Función principal para construir y enviar una consulta a la API de OpenAlex. Permite especificar la entidad (*works*, *authors*, *institutions*, etc.) y parámetros personalizados.
	oa_get()	Envía la consulta generada por `oa_query()` y devuelve los datos en formato de lista o `data.frame`.
	oa_entity()	Acceso directo a una entidad específica mediante su identificador OpenAlex. Sería para recuperar un solo registro completo.
	oa_pagination()	Permite recuperar múltiples páginas de resultados automáticamente, útil para colecciones grandes.
Gestión de filtros	*oa_filters()*	Visualiza los filtros disponibles para una entidad. Revísese los filtros que ofrece para una mejor búsqueda, filtros como por ejemplo: concepts.id, from_publication_date, etc.
	oa_make_filter()	Crea filtros compuestos fácilmente combinando múltiples condiciones para la consulta.
Conversión de resultados	*oa_tidy()*	Convierte los resultados en formato anidado de la API a un `tibble`/`data.frame` plano, listo para el análisis en `R`.
Integración con R	Con `dplyr`, `igraph`	Se puede integrar/manejar con herramientas comunes para análisis, redes, visualización o modelado de datos.

Otros aspectos a tener en cuenta serían que su obertura es limitada en comparación con bases comerciales como WoS, etc. Tiene dependencia de la API de OpenAlex, por lo que el tiempo de consulta puede variar según la consulta y/o el servidor. También se debe tener algo de experiencia en `R`.

1.7.2. rcrossref

Se menciona en Chamberlain et al. (2014), que rcrossref es una librería que permite interactuar de manera directa con la API de Crossref, proporcionando acceso a los datos bibliográficos de las publicaciones. Como se vio en la Subsección 1.5.5, Crossref es una organización que mantiene una base de datos de identificadores DOI, lo que facilita la búsqueda, identificación y citación de los trabajos académicos.

Características principales

Entre las diferentes características que ofrece la librería `rcrossref`, vease la Tabla 1.4 y las expuestas:

- Realizar búsquedas en la base de datos de Crossref, mediante consultas detalladas, permitiendo ajustar los filtros según las necesidades del usuario (título, autor, ISSN-L, DOI, etc.). Permitiendo así la recuperación de datos asociados a publicaciones, como citas, autores, afiliaciones, y fechas de publicación.
- Explorar tendencias en áreas específicas mediante consultas avanzadas, y/o combinación de filtros.

- Una vez conectadas/descargadas las BBDD, se pueden combinar con herramientas como `bibliometrix` , y se podrían: calcular índices bibliométricos, análisis de co-autoría, co-citación, análisis de co-palabras, clústeres, también se podría un análisis temático y detección de tendencias.

Tabla 1.4: Funciones principales del paquete `rcrossref`

Enfoque	Función	Descripción
Búsqueda general	*cr_works(dois = ...)*	Consulta trabajos (artículos, libros, capítulos, etc.) usando filtros como DOI, título, autor, año, etc.
	cr_members()	Devuelve información sobre los miembros registrados de Crossref (editoriales, instituciones).
	cr_funders()	Permite recuperar datos sobre agencias financiadoras y vincularlas a obras financiadas.
Filtrado y metadatos	*cr_filter()*	Utiliza filtros predefinidos (por ejemplo: tipo de publicación, fecha, editor, ISSN, etc.) para personalizar las búsquedas.
	cr_types()	Obtiene distintos tipos de registros que reconoce Crossref (journal-article, proceedings, etc.).
	cr_licenses()	Devuelve información sobre licencias de acceso abierto asociadas a los documentos.
Exportación de resultados	Integración con `tibble`, `dplyr`, `jsonlite`	Los resultados se pueden procesar y analizar fácilmente usando herramientas de tidyverse o convertir desde JSON.

Otros aspectos a tener en cuenta serían que su obertura es limitada en comparación con bases comerciales como WoS. Puede ocurrir que, para ciertas consultas, existan áreas o revistas que no estén presentes. La API de Crossref impone restricciones en la cantidad de solicitudes, lo que puede dificultar proyectos de gran escala.

Las herramientas mostradas anteriormente (`openalexR` y `rcrossref`), son solo gestores de conexión, búsqueda y descarga, no permitirían un procesado de estos. Las herramientas presentadas a continuación sí lo permitirían.

1.7.3. Bibliometrix

`Bibliometrix` es un paquete de `R` orientado a la investigación cuantitativa en diversos campos de la cienciometría y bibliometría, es un paquete amplio con varias funciones interesantes. En Aria et al. (2024), se mencionan las funciones que incluye y algunos ejemplos, a continuación se describen sus características principales.

Características y funciones principales en Bibliometrix

En la documentación oficial de la librería, se puede encontrar con mayor detalle las funciones, filtros, y ejemplos que incluye (ver Tabla 1.5).

- Aunque permite manejar datos de repositorios comerciales (Scopus, WoS, etc.), se manejarán las de acceso libre (OpenAlex y/o Crossref).
- Permite hacer análisis bibliométricos básicos y avanzados como: cálculo de índices bibliométricos, número de publicaciones, citas, índice h, índice g, etc.), descritas en la Sección 1.6.

- Análisis de redes de coautoría, co-citación, co-palabras, etc. Para usuarios que no manejan programación o `R`, la interfaz gráfica (biblioshiny) es de gran utilidad.
- Visualización avanzada (mapas científicos, gráficos de redes, análisis temático y detección de tendencias etc.).

Tabla 1.5: Funciones principales del paquete `bibliometrix()`

Enfoque	**Función**	**Descripción**
Conversión de datos	*Convert2df()*	Crea un dataframe cuyas columnas utilizan la codificación estándar de WoS.
Análisis de datos bibliométricos	*biblioAnalysis()*	Calcula las principales medidas bibliométricas y devuelve un objeto de tipo `bibliometrix`.
	Summary()	Resume los principales resultados del análisis bibliométrico, mostrándolos en tablas.
	Plot()	Visualiza gráficos básicos de índices calculados.
	citations()	Identifica las referencias o autores más citados.
	localCitations()	Calcula citas locales de autores y documentos de una colección bibliográfica.
	dominance()	Devuelve un Dataframe que incluye los k autores más productivos y variables típicas de un análisis de dominancia.
Índices bibliométricos	*Hindex()*	Mide la productividad e impacto (*Índice H*).
	lotka()	Índice Lotka para la productividad científica.
Extracción de términos	*termExtraction()*	Extrae términos de resúmenes, títulos, palabras clave, etc.), de una colección bibliográfica.
Redes bipartidistas	*cocMatrix()*	Otorga una red bipartita que puede transformarse en una colección de redes como acoplamiento, co-citación, etc.
Matriz de similitud	*biblioNetwork()*	Calcula las redes de acoplamiento más utilizadas.

Se debe considerar que el manejo de grandes conjuntos de datos requiere algo de tiempo y/o recursos computacionales para su procesamiento, por lo que estos deberían tener una inspección o tratamiento previo para depurar inconsistencias.

1.7.4. Scimetr

Esta librería es una propuesta reciente y similar a `bibliometrix()`, está diseñada para ser una herramienta ligera, flexible, y de fácil uso para usuarios de `R`. En Fernandez-Casal et al. (2023), su documentación y/o web oficial, se menciona que la librería incluye herramientas orientadas hacia la investigación cuantitativa en los campos de la bibliometría y cienciometría. Permitiendo la conexión, descarga y transformación de datos bibliométricos provenientes de WoS. Sus funciones de cálculo permitirían realizar análisis básicos y avanzados, como por ejemplo: calcular la producción científica, el impacto de publicaciones, y la obtención de índices

bibliométricos (índice h e índice g). Para posteriormente elaborar sumarios de los datos o una representación gráfica.

Características y funciones principales en Scimetr

En la documentación oficial de la librería, se puede encontrar con mayor detalle las funciones, filtros, y ejemplos que incluye. Se describirá de manera rápida las funciones que se han usado, ver Tabla 1.6.

- Permite hacer análisis bibliométricos básicos: cálculo de índices bibliométricos (número de publicaciones, citas, etc.), Sección 1.6, entre otros.
- Integración con otros paquetes de `R`, como `tidyverse` y `ggplot2`, etc.
- Simplicidad y facilidad de uso, al poseer referencias y ejemplos, que dan a conocer el contenido más fácil.
- Representación y visualización de los datos bibliométricos.

Tabla 1.6: Funciones principales del paquete `Scimetr`

Enfoque	**Función**	**Descripción**
Importación de datos (WoS)	*ImportSources.wos()*	Permite leer e importar archivos bibliométricos delimitados por tabulaciones codificados en UTF-8, generando un dataframe
Conversión de datos	*CreateDB.wos()*	Crea una BBDD relacional, con múltiples tablas, cuyas columnas utilizan la codificación estándar de WoS.
Análisis de datos bibliométricos	*summary_year()*	Resume los principales resultados del análisis bibliométrico, mostrándolos en tablas por años.
	plot.wos.db()	Representa gráficamente la información relevante, producidos por los objetos *wos.db*.
Índices bibliométricos	*TC.authors()*	Muestra la productividad e impacto de los autores (*Índice h, Índice g*).
Filtros `get.id<Tabla>()`	*get.idAuthors()*	Permite buscar los id de los autores, dentro de la base de datos.
	get.idAreas()	Devuelve códigos de las áreas de investigación (matemáticas, informática, etc.).
	get.idDocs()	Devuelve índices de los trabajos correspondientes a otro filtro, por ejemplo los trabajos de *get.idAuthors()*, y luego obtener información de ese ente con *base::summary()* o *summary_year()*.

Actualmente solo admite datos en formato estándar Wos (Tabla 2.1), se podría utilizar la librería si se hiciese una transformación (reetiquetado) a los campos de la matriz devuelta por la consulta de las librerías `openalexR` y `rcrossref`. Sin embargo, puede ocurrir que las matrices devueltas por las librerías, no contengan las columnas o campos necesarios para poder modificarlas, esto a causa de la propia API, consulta o del repositorio, haciendo que esto sea más extenso o no sea posible por falta de datos.

Capítulo 2

Diseño e implementación

2.1. Descripción del caso de estudio

En esta sección se dará una pequeña descripción referente al caso de estudio, el cual se centrará en el *Sistema Universitario de Galicia* (SUG), al cual se aplicarán las distintas herramientas bibliométricas y permitirán identificar las tendencias y patrones en su producción científica, destacando áreas de fortaleza.

2.1.1. Sistema Universitario de Galicia

Se entiende que el SUG, es una organización pública encargada del desarrollo de la educación superior en Galicia, y está conformada por las tres universidades públicas gallegas: *Universidad de Santiago de Compostela* (USC), *Universidad de A Coruña* (UDC) y la *Universidad de Vigo* (UVigo) (ver Touriñan-López (2025)). El caso de estudio se centra en el análisis bibliométrico de la producción científica del SUG, con la finalidad de evaluar la productividad científica, el impacto de las publicaciones, entre otros, véase la Sección 1.2 (objetivos)

2.1.2. Metodología aplicada al caso de estudio

Los datos se obtendrán de los repositorios abiertos de OpenAlex y Crossref, mediante las librerías `openalexR` y `rcrossref`, que como se vio anteriormente, permiten hacer consultas y/o descargas a través de la interacción de sus API. Se considerarán además los datos ofrecidos por WoS para poder hacer una comparación sistemática u ordenada. Y se procesarán principalmente con Bibliometrix. Se seleccionarán publicaciones asociadas a las universidades en los últimos diez años (2014-2024), para ello será necesario hacer la conexión desde `R` o descargar los metadatos desde los repositorios.

Mediante esta formulación se podrá conocer detalles como:

- La productividad científica de las universidades del SUG.
- El impacto de las publicaciones.
- Patrones de colaboración institucional y de autoría (si existen).
- Áreas temáticas de relevancia o evolución de contenido, que destacan en la producción científica.
- Estado de los autores (indices,citaciones, etc)

Lo anterior solo es un esquema o planificación general, que tiene intención de otorgar una idea de la estructura del área de estudio. La obtención de datos, conexión, tratamiento, entre otros procedimientos, se explicarán posteriormente en las siguientes secciones.

2.2. Obtención y preparación de los datos

En esta sección se muestra una visón general del proceso práctico, que se ha llevado a cabo, para la obtención de los datos de los diferentes repositorios de datos abiertos analizados, los parámetros de búsqueda general serán:

- Instituciones asociadas al SUG (USC,UDC UVigo).
- Periodo temporal: 2014-2024.
- Tipo de documentos: Artículos científicos, capítulos de libros y actas de congresos, revisiones.

Lo anterior sería un esquema simple de los elementos que se necesitarían para poder aplicar las funciones de búsqueda: `OpenaAlex::oa_query()` o `rcrossref::cr works()`, las descripciones anteriores poseen atributos dentro de las consultas, estos se explicarán posteriormente.

2.2.1. Obtención de datos desde OpenAlex

Como se mencionó previamente (Subsección 1.5.4), OpenAlex es un repositorio abierto que ofrece información bibliográfica. En lo que sigue, se presenta cómo interactuar y conectar con la API de OpenAlex utilizando el lenguaje de programación `R`. Una ventaja que tiene OpenAlex y que evita el uso excesivo de filtros o BBDD, es que al tener entidades relacionadas, no hace falta descargar una base de datos para cada entidad, ya que, al descargar una sola entidad esta estará relacionada con otras, por ejemplo, en este apartado se mostrarán todos los trabajos o publicaciones (libros, artículos, revisiones, etc.) que hay en el periodo de estudio. El detalle de conexión podría ser el siguiente:

- **Instalación y carga de la librería**
  - Se debe cargar la librería `openalexR`. Si no se tiene instalada, descárguese con la consola de `R`, o con el comando `install.packages()`.
- **Definir la consulta**
  - **Institución:** Identificar las universidades de interés, en este caso, las pertenecientes al SUG. Establézcase sus identificadores únicos en (ROR ID) o (`institution.id` en OpenAlex), para más información véase en la sección de filtros para entidades en la página web oficial de OpenAlex. Para poder hallar los identificadores se puede consultar en la página de Ror.org, Wikidata o directamente en la página oficial de OpenaAlex. Para ello, se debería buscar la universidad de interés y acceder a una publicación, luego dar en la opción `View api`, y finalmente buscar aquellos autores suscritos a la universidad de interés. Ejemplo, para la USC, el RorID sería `"030eybx10"`, véase Figura 2.1.
  - **Periodo de tiempo**: Se establece mediante los parámetros `from_publication_date = "2014-01-01"` y `to_publication_date = "2024-12-31"`.
- **Obtener datos de publicaciones**:
  - Obténgase mediante la combinación de funciones como: `oa_query()`, `oa_request()` y `oa2df()`.
  - `oa_query()`: Permite establecer los parámetros de búsqueda vistos anteriormente, ejemplo: fechas, institución, entre otros.
  - `oa_request()`: Permite pasar los parámetros de búsqueda vistos anteriormente a la API de OpenAlex, esto devolverá un objeto lista.
  - `oa2df()`: Adapta la lista anterior a un formato matriz y con `oa2bibliometrix()` (de `Bibliometrix`), se adapta esta matriz a un formato similar al de WoS.

Todo lo anterior puede entenderse y verse en el siguiente código. La matriz de interés que se trabajará será M, similar a la notación de `Bibliometrix()`.

Fragmento de código 2.1: Ejemplo de consulta para Works(Obras) con `openalexR` para la USC, devuelve una matriz de colección bibliográfica

```
# Ejemplo de consulta para Works(Obras) con openalexR (USC)
    Id_openalex="I200284239" ; Ror_Id ="030eybx10"

    query <- oa_query(entity = "works",
    from_publication_date = "2014-01-01", to_publication_date = "2024-12-31",
    verbose = TRUE,
    authorships.institutions.id=Id_openalex, authorships.institutions.ror=Ror_Id )

    res <- oa_request(query_url = query, count_only = FALSE, verbose = FALSE )
    df <- oa2df(res, entity = "works")
    M  <-  oa2bibliometrix(df);
```

Figura 2.1: Ejemplo de consulta para la obtención del ROR ID para la USC en OpenAlex.

2.2.2. Obtención de datos desde Crossref

`Rcrossref` es otra librería de `R` que permite acceder a datos bibliográficos a través de la API de Crossref. A diferencia de OpenAlex, Crossref es una base de datos que indexa datos de publicaciones, incluyendo DOIs, títulos, autores, revistas y más (Subsección 1.5.5). El detalle de conexión podría ser el siguiente :

- **Instalación y carga de la librería**: Se debe cargar en memoria la librería `rcrossref`. Si no se tiene instalada, descárguese con el comando `install.packages()` o con la consola de `R`.
- **Definir la consulta**
  - **Institución:** Identificar las universidades de interés, en este caso, las pertenecientes al SUG. En Crossref, no se utilizan IDs de instituciones como en OpenAlex, sino que se filtra por el nombre de la institución o afiliación. Esta es una desventaja, se deberían considerar las variaciones en nombre de las universidades (ver Figura 2.2). Si se usa `funder=Funder_id`, como unicódigo, se limitara a pocos resultados.

University of Santiago de Compostela (Q1473677)

Item Discussion Read View history Tools

university in Spain

USC | Universidad de Santiago de Compostela | Universidade de Santiago de Compostela | Universidad Santiago de Compostela | usc.es

In more languages

Configure

Language	Label	Description	Also known as
default for all languages	No label defined	–	
English	University of Santiago de Compostela	university in Spain	USC Universidad de Santiago de Co… Universidade de Santiago de C… Universidad Santiago de Comp… usc.es
Spanish	Universidad de Santiago de Compostela	universidad pública española	Universidade de Santiago de C… USC
Catalan	Universitat de Santiago de Compostel·la	No description defined	USC Universitat de Santiago de Com…
Galician	Universidade de Santiago de Compostela	universidade pública galega con sede en Santiago de Compostela	USC Universidade de Santiago Universidade de Compostela

All entered languages

Figura 2.2: Ejemplo búsqueda identificadores de la USC en Wikidata.

- **Periodo de tiempo**: Filtrar publicaciones (intervalo), parámetros `from_pub_date = "2014-01-01"` y `until_publication_date = "2024-12-31"`.

- **Obtener datos de publicaciones**: Obténgase mediante `cr_works()`, combinando los campos anteriores.

Fragmento de código 2.2: Ejemplo de consulta para Works(Obras) con `rcrossref` para la USC.

```
names_usc <- c("Universidade de Santiago de Compostela","Univ. Santiago de Compostela",
"University of Santiago de Compostela","USC","Univ Santiago de Compostela",
"Santiago de Compostela University")
all_results <- list(); limit <- 950 ; cursor <- "*" ;
while (!is.null(cursor)) {
  tryCatch({ result <- cr_works( filter = c(
     from_pub_date = "2014-01-01", until_pub_date = "2024-12-31"),
      query = paste0(names_usc, collapse = " OR "), cursor=cursor,limit=limit)
    if (!is.null(result$data)) {
      all_results <- append(all_results, list(result$data))}
    cursor <- result$meta$`next-cursor`
    if (is.null(cursor)) {break}
    Sys.sleep(1)}, error = function(e) {
    cat("Se produjo un error: ", e$message, "\n");cursor <- NULL }) }
results <- do.call(rbind, all_results)
works_usc_crf <- results%>%
  filter(purrr::map_lgl(author, ~ is.data.frame(.x) && nrow(.x) > 0))
```

2.2.3. Obtención de datos desde WoS

Los datos del repositorio de WoS son de acceso limitado para el público en general, para poder descargarlos se deberá iniciar sesión para el acceso a su sistema y/o API. En este caso, mediante una cuenta institucional se procedió a dicha ejecución. Se detallarán rápidamente los pasos necesarios para la obtención de las publicaciones. Una explicación más detallada puede encontrarse en lo expuesto por Fernandez-Casal et al. (2023), o para un modelado de conexión más avanzado véase Fernández-Ramos (2021). Una vez accedida a la API de WoS, modifíquese los siguientes campos de interés:

- **Definir la consulta**
  - **Institución:** Se agregará el nombre de la institución u afiliación, mediante `OG=[Affiliation]`. Por ejemplo, para consultar las publicaciones de la USC, se puede utilizar: `OG = (Universidade de Santiago de Compostela)`. El buscador de WoS es interactivo y autocompleta conforme se escribe.
  - **Periodo de tiempo**: Se establece mediante el filtro `Publication Date::Custom`, indicando las fechas de inicio y fin en formato YYYY-MM-DD.
  - **Colecciones a incluir**: En el caso de esta investigación se consideraron únicamente las bases SCI, SSCI y A&HCI, ya que representan las colecciones principales de WoS orientadas a ciencia, ciencias sociales y humanidades, respectivamente. Estos filtros se ilustran en la Figura 2.3.
- **Descargar de resultados:** WoS permite exportar los registros en varios formatos: RefWorks, RIS, Excel, BibTeX, TxT, entre otros. En este caso, se seleccionó el formato TXT, el cual es compatible con `R`. La descarga debe realizarse en grupos de hasta 500 registros, activando la opción `Full Records and Cited References` para capturar la mayor cantidad de variables o metadatos disponibles.
- **Importación al entorno de R:**: Una vez descargados todos los registros, se importan los registros para unificarlos y poder trabajar con ellos.

Fragmento de código 2.3: Permite importar los TXT y generar una matriz apta para trabajarlos, esto con la ayuda de `bibliomterix::convert2df()`

```
import <- function(directorio) {
  archivos <- list.files(directorio,pattern="\\.txt$",full.names=TRUE)
  # Convertir cada archivo a un dataframe
  lista_df <- lapply(archivos, function(directorioe) {
    convert2df(directorioe, dbsource = "wos", format = "plaintext") })
  # Campos/columnas en común
  columnas_comunes <- Reduce(intersect, lapply(lista_df, names))
  lista_directoriotrada <- lapply(lista_df,function(df)df[,columnas_comunes])
  return(do.call(rbind, lista_directoriotrada)) }

usc_M_wos <- import(directorio = "C:\\Users\\...")
```

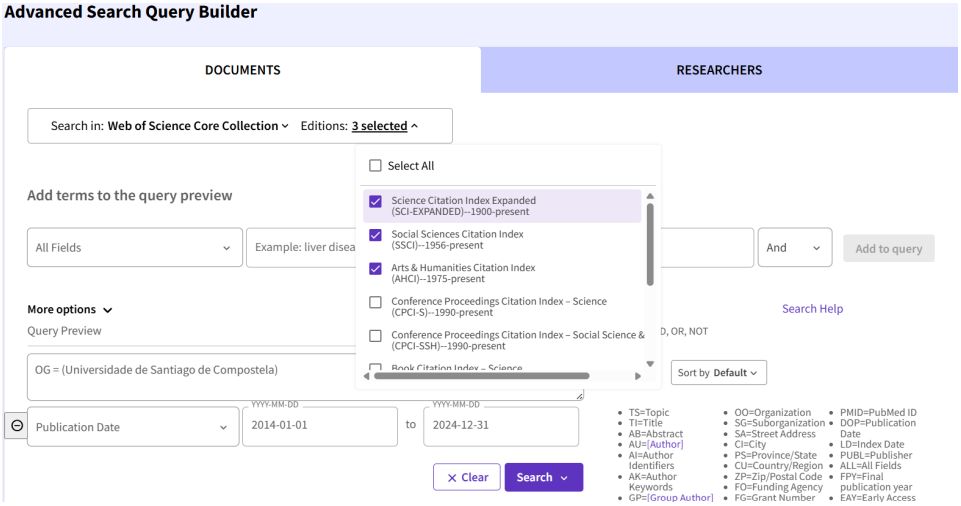

Figura 2.3: Ejemplo de búsqueda de publicaciones mediante la API de WoS

2.3. Procesamiento de datos descargados

La informacion obtenida desde OpenAlex o Crossref, incluyen datos complejos (anidados), por lo que es necesario un proceso de preselección, limpieza y preparación antes del análisis. Este proceso incluiría:

- **Transformación de datos en formato R (DataFrame)**: Los datos se deben convertir en un data frame. En OpenAlex, este paso fue realizado gracias a `oa2df()`. Por su parte, `rcrossref` devuelve ya una matriz, solo se debería hacer una transformación al obtener un formato JSON u otro.
- **Reetiquetado o normalización de campos clave**: Uno de los procesos más importantes, se debe homogeneizar los nombres de las variables para poder adecuarlas a Bibliometrix o Scimetr. Para los datos obtenidos de OpenAlex usar `oa2df()`.

2.3.1. Etiquetado para openalexR y rcrossref

Para el reetiquetado de datos es importante tener en cuenta los acrónimos para variables que tiene Wos, estos acrónimos deberán estar en las BBDD que se han descargado. Los paquetes `bibliometrix` y `Scimetr` aceptan estos acrónimos, véase una expansión de estos en WoS Tags o Fernandez-Casal et al. (2023). Los paquetes `openalexR` o `rcrossref` devuelven una matriz de datos. No obstante, debido a la falta de una estructura categórica estandarizada, sería necesario realizar un proceso de transformación para adaptarlos al formato compatible con la librería `bibliometrix()`, ver Tabla 2.1. Con `openalexR` no sería necesaria la estandarización de campos, mediante `oa2bibliometrix(oa2df())`, se tendría una matriz en formato WoS. En el caso de `rcrossref` se necesitarán funciones auxiliares que permitan deslistar los datos anidados que la consulta `cr_works()` devuelva. Estas funciones están disponibles en el Apéndice 4

2.4. Variables del análisis de correspondencias

Para explorar la asociación entre universidades y el tipo de producción que generan, se aplicará un análisis de correspondencias (AC) sobre una tabla de contingencia construida por las variables DT (tipo de documento) y SUG, que representa a las tres universidades gallegas analizadas: UDC, USC y UVigo. Como se mencionó en la Subsección 1.6.5, permitirá ver las relaciones entre categorías cualitativas en un espacio reducido, facilitando la interpretación de patrones de asociación entre instituciones y sus orientaciones hacia la producción. Se necesitará introducir primero las variables a relacionar junto con sus descripciones. En la página web oficial de OpenAlex y WoS, se entiende que hay publicaciones de distinto tipo, en base a ello se decidió reagruparlas para poder tener una representación más categorizada y no multivariada de las mismas etiquetas. Véase la Tabla 2.2 y la descripción del siguiente código:

Fragmento de código 2.4: Realiza un reetiquetado de las categorías existentes en la variable DT (tipo de documento), para realizar el análisis de correspondencias. Repositorio de OpenAlex

```
# Ventana/Script OpenAlex

load("Udc_M_Opnlx.Rda"); load("Usc_M_Opnlx.Rda"); load("Udv_M_Opnlx.Rda")
data_sug_opnlx <- bind_rows(Udc_M_Opnlx,Usc_M_Opnlx,Udv_M_Opnlx)
data_sug_opnlx$SUG <- factor(rep(c(1,2,3), times = c(nrow(Udc_M_Opnlx),
                        nrow(Usc_M_Opnlx), nrow(Udv_M_Opnlx))))
levels(data_sug_opnlx$SUG)=c("1"="UDC","2"="USC","3"="UDV")
data_sug_opnlx  <- data_sug_opnlx %>% filter(PY %in% 2014:2024)

# Agrupacion-Etiquetación en OpenAlex
data_sug_opnlx$DT_aux <- dplyr::case_when(
      data_sug_opnlx$DT %in% c("ARTICLE", "REVIEW") ~ "Artículo o revisión",
      data_sug_opnlx$DT %in% c("BOOK", "BOOK-CHAPTER") ~ "Libro o Capítulo",
      data_sug_opnlx$DT %in% c("EDITORIAL", "LETTER", "ERRATUM", "REPORT","PARATEXT",
    "REFERENCE-ENTRY", "DISSERTATION", "OTHER") ~ "No evaluable",
      data_sug_opnlx$DT %in% c("DATASET", "PREPRINT") ~ "Difusión sin revisión",
      TRUE ~ "Otro")
```

Tabla 2.1: Comparativa de reetiquetado entre rcrossref y openalexR

Variable WoS	Descripción	rcrossref	openalexR
TI	Título del trabajo	`title`	`title`
AU	Autores (separados por ';')	`author` (procesar para extraer)	`author` (extraer)
SO	Fuente (nombre de la revista)	`container.title`	`so`
PY	Año de publicación	`created`, o `issued`	`publication_year`
DI	DOI (identificador del trabajo)	`doi`	`doi`
VL	Volumen de la revista	`volume`	`volume`
IS	Número de la edición (issue)	`issue`	`issue`
PP	Páginas (página inicial y final)	`page`	`first_page`, `last_page`
AB	Resumen (abstract)	`abstract`	`ab`
CR	Referencias (separadas por ';')	`references` (extraer)	`referenced_works`
TC	Número de citas	`is.referenced.by.count`	`cited_by_count`
UT	Identificador único	`alternative_id` o `doi`	`id` de OpenAlex
DB	Base de datos de origen	Asignar manualmente, ejemplo: CrossRef, Scopus, WoS, etc.	Asignar: CrossRef, Scopus, WoS, etc.

Fragmento de código 2.5: Realiza un reetiquetado de las categorías existentes en la variable DT (tipo de documento), para realizar el análisis de correspondencias. Repositorio de WoS

```
# Ventana/Script WoS

load("udc_M_wos.Rda");load("usc_M_wos.Rda"); load("udv_M_wos.Rda")
data_sug_wos <- bind_rows(udc_M_wos,usc_M_wos,udv_M_wos)
data_sug_wos$SUG <- factor(rep(c(1,2,3), times = c(nrow(udc_M_wos),
                        nrow(usc_M_wos),nrow(udv_M_wos))))
levels(data_sug_wos$SUG)=c("1"="UDC","2"="USC","3"="UDV")

# Agrupacion-Etiquetación en WoS
data_sug_wos <- data_sug_wos %>% filter(PY %in% 2014:2024) %>% mutate(
      DT_aux = case_when(
        # Articulos-Review(y variantes)
        DT %in% c("ARTICLE", "ARTICLE; BOOK CHAPTER",
        "ARTICLE; DATA PAPER","ARTICLE; EARLY ACCESS", "ARTICLE; PROCEEDINGS PAPER",
        "ARTICLE; PUBLICATION WITH EXPRESSION OF CONCERN",
        "ARTICLE; RETRACTED PUBLICATION",
        "ARTICLE; WITHDRAWN PUBLICATION","REVIEW", "REVIEW; BOOK CHAPTER",
        "REVIEW; EARLY ACCESS","REVIEW; RETRACTED PUBLICATION")~"Artículo o revisión",

        # Libros y capítulos (variantes de BOOK)
        DT %in% c("BOOK","BOOK CHAPTER","BOOK REVIEW",
                  "BOOK REVIEW; EARLY ACCESS")~"Libro o Capítulo",

        # Material editorial y cartas
        DT %in% c("EDITORIAL MATERIAL", "EDITORIAL MATERIAL; BOOK CHAPTER",
        "EDITORIAL MATERIAL; EARLY ACCESS","LETTER",
              "LETTER; EARLY ACCESS")~ "No evaluable",

        # Correcciones y retracciones
        DT %in% c("CORRECTION","CORRECTION; EARLY ACCESS",
                  "RETRACTION")~"No evaluable",

        # Difusión sin revisión por pares
        DT %in% c("MEETING ABSTRACT","PREPRINT","DATASET")~"Difusión sin revisión",

        # Otros tipos diversos
        DT %in% c("NEWS ITEM","BIBLIOGRAPHY","BIOGRAPHICAL-ITEM",
                  "SOFTWARE REVIEW")~"Otro",
        # Casos especiales
        DT %in% c("EXHIBIT REVIEW","POETRY","REPRINT") ~ "Otro",
        # Por defecto
          TRUE ~ "Otro"))
```

Tabla 2.2: Agrupación considerada para los tipos de documentos en WoS y OpenAlex. Los asteriscos (*) indican variantes compuestas (ej. ARTICLE; BOOK CHAPTER, ARTICLE; RETRACTED PUBLICATION, REVIEW; BOOK CHAPTER)

Agrupación	**Etiqueta WoS (DT)**	**Etiqueta OpenAlex (DT)**	**Descripción**
Artículo o revisión	ARTICLE*, REVIEW*	ARTICLE, REVIEW	Son publicaciones principales evaluadas por pares. Es decir, artículos científicos y revisiones.
Libro o Capítulo	BOOK, BOOK CHAPTER, BOOK REVIEW	BOOK, BOOK-CHAPTER	Se contemplan los libros académicos o capítulos incluidos en obras colectivas.
No evaluable	EDITORIAL MATERIAL*	EDITORIAL, LETTER, ERRATUM, ...	Documentos no sujetos a revisión por pares, o con valor académico y/o científico limitado.
Difusión sin revisión	MEETING ABSTRACT, PREPRINT, DATASET	DATASET, PREPRINT	Material de circulación temprana o sin revisión formal (repositorios, conjuntos de datos).
Otro	EXHIBIT REVIEW, POETRY, REPRINT	other	Elementos marginales o no científicos, incluidos ítems de divulgación o creativos.

Capítulo 3

Resultados: caso de estudio del SUG

En este capítulo se aplicarán técnicas bibliométricas a los datos obtenidos de OpenAlex y Crossref, relacionados con la escala e impacto de la investigación de las universidades del SUG. Para ello, se utilizarán herramientas o librerías en R, como el paquete `bibliometrix`, que permite analizar la cantidad, impacto y grado de internacionalización, entre otros, de la producción científica. A continuación, se describen los análisis realizados:

3.1. Productividad científica de las universidades del SUG.

La productividad científica se mide a través del número total de publicaciones generadas por cada universidad en el período 2014-2024. Es decir, en base a las métricas vistas en la Sección 1.6, se calcula el total de publicaciones por institución y la distribución anual de publicaciones. En concreto, para la estimación de la producción científica de las universidades se llevan a cabo las siguientes tareas.

- Cálculo del número de publicaciones por universidad y año.
- Estimación de la evolución de la evolución de la producción científica a lo largo del tiempo.
- Comparación entre las universidades en términos de producción.

3.1.1. Número de publicaciones por universidad y año.

Los resultados obtenidos se presentan en la Tabla 3.1, así como en las figuras Figura 3.1 y Figura 3.2. A partir de estos resultados, y tomando como referencia los datos de WoS, se observa que el repositorio de OpenAlex contiene un mayor número de registros de publicaciones en comparación con Crossref y WoS. Esta diferencia era esperable debido a que OpenAlex pretende ser una BBDD exhaustiva, mientras WoS, por su parte, indexa solo a revistas que se ajustan a unos exigentes parámetros de calidad.

Con el fin de facilitar la comparación entre repositorios, el número de publicaciones se representó en escala logarítmica, como $log(N_P)$, debido a que las escalas de OpenAlex y WoS son notablemente mayores, lo que podría opacar los resultados de CrossRef. Los valores originales (sin transformación), se encuentran en la Tabla 3.1. En cuanto a la producción académica, la USC destaca como la universidad con mayor número total de publicaciones, seguida por la UVigo, siendo la UDC la tercera universidad por producción científica. Sin embargo, los resultados obtenidos desde Crossref podrían estar sesgados; de hecho, es posible que los registros en estas BBDD sean incompletos o no estén actualizados, a diferencia de lo que ocurre en WoS u OpenAlex. En estos resultados, también puede influir el funcionamiento de la API de CrossRef, accedida mediante la librería `rcrossref`. Finalmente, se observa una tendencia creciente en el número de publicaciones a lo largo del tiempo en los tres repositorios. Particularmente, el periodo comprendido entre los años 2018 y 2022 representa un punto de inflexión, en el que se evidencia un aumento significativo en la producción científica de las universidades del SUG, asumiendo la información que proporcionan OpenAlex, Crossref y WoS.

Fragmento de código 3.1: Calcula el número de publicaciones (N_P) para cada universidad del SUG, devolviendo un formato tabla. Posteriormente, grafica el (N_P) a través del tiempo. Resultados para los repositorios de OpenAlex, Crossref y WoS, respectivamente.

```
# Número de publicaciones por año
productividad_anual_wos <- data_sug_wos %>%mutate(Fuente = "WoS") %>%
  group_by(SUG,PY,Fuente) %>% summarise(total_publicaciones = n())
productividad_anual_crsrf  <- data_sug_crf %>% mutate(Fuente="CrossRef")%>%
  group_by(SUG,PY,Fuente)%>%summarise(total_publicaciones=n(),.groups="drop")
productividad_anual_opnlx <- data_sug_opnlx %>%mutate(Fuente = "OpenAlex") %>%
  group_by(SUG,PY,Fuente) %>%summarise(total_publicaciones = n())
# Ejemplo del gráfico OpenAlex-Wos
Mee <- bind_rows(productividad_anual_wos, productividad_anual_opnlx)
ggplot(Mee, aes(x = factor(PY), y = log(total_publicaciones),
    color = SUG, linetype = Fuente, group = interaction(SUG, Fuente))) +
    geom_line()+geom_point(size = 1.5)+labs(x = "Año",y="log(Nº.Publicaciones)",
    color="SUG",linetype="Fuente")+theme_minimal()+theme(legend.position="right")
# Ejemplo del gráfico CrossRef-Wos
Mee <- bind_rows(productividad_anual_wos, productividad_anual_crsrf)
ggplot(Mee, aes(x = factor(PY), y = log(total_publicaciones),
    color = SUG, linetype = Fuente, group = interaction(SUG, Fuente))) +
    geom_line()+geom_point(size = 1.5)+labs(x = "Año",y="log(Nº.Publicaciones)",
    color="SUG",linetype="Fuente")+theme_minimal()+theme(legend.position="right")
```

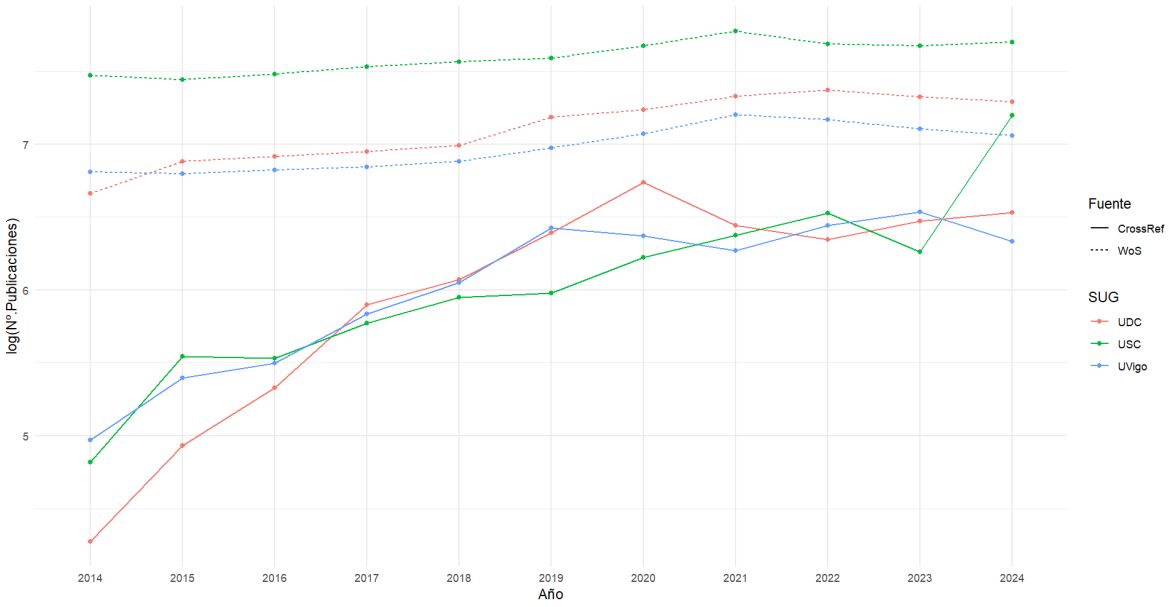

Figura 3.1: Evolución, productividad científica SUG, CrossRef.

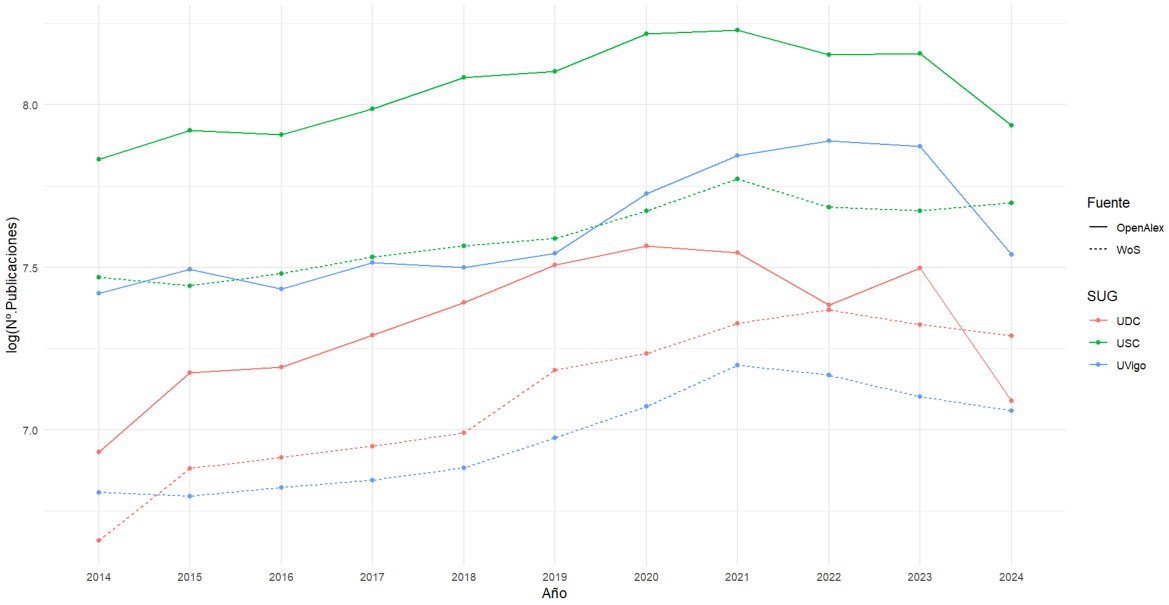

Figura 3.2: Evolución, productividad científica SUG, OpenAlex.

Tabla 3.1: Número de publicaciones por universidad, año y fuente.

Año	Crossref			OpenAlex			WoS		
	UDC	**USC**	**UVigo**	**UDC**	**USC**	**UVigo**	**UDC**	**USC**	**UVigo**
2014	71	124	143	1023	2523	1670	780	1753	904
2015	139	255	222	1308	2754	1796	972	1707	893
2016	208	253	245	1329	2720	1691	1006	1772	918
2017	361	322	339	1466	2946	1835	1041	1866	938
2018	432	383	420	1621	3243	1806	1086	1930	975
2019	602	394	618	1819	3305	1887	1318	1976	1070
2020	836	505	596	1929	3709	2267	1387	2151	1177
2021	623	588	528	1889	3749	2550	1520	2375	1338
2022	574	685	623	1609	3478	2670	1586	2175	1297
2023	647	523	687	1803	3493	2623	1515	2149	1215
2024	690	1332	562	1199	2795	1880	1465	2204	1163
Total	5183	5364	4983	16995	34715	22675	13676	22058	11888

3.2. Impacto de las publicaciones

El impacto de las publicaciones se mide principalmente a través del cálculo de métricas de citación (índice h, índice g, etc.). Los resultados principales se muestran en las tablas referenciadas dentro de cada apartado. Además de las tablas, se han contruido una serie de gráficos, a partir de la misma información, cuyo objeto es presentar la información de una forma más intuitiva, rápida e ilustrativa, además de poner el foco en las tendencias temporales.

Específicamente, en la presente sección, se han llevado a cabo las siguientes tareas para medir el impacto de la producción científica de las tres universidades del SUG:

- Cálculo del número total de citas por universidad y por año.
- Calculo del promedio de citas por publicación.
- Obtención del índice h de cada universidad.

3.2.1. Número total de citas por universidad y por año.

La Figura 3.3 y la Figura 3.4 proporcionan, de forma intuitiva, la información contenida en las tablas Tabla 3.3 y Tabla 3.2, respectivamente. Estas tablas reúnen información general sobre diversas métricas del SUG por año, fuente de datos, y universidad. Con el fin de facilitar la comparación entre las distintas fuentes, en las representaciones gráficas se empleó la transformación $log(Total\ Citas)$, mientras que en las tablas se presentan los valores originales sin transformar. A partir del análisis de las citaciones según Crossref, se observa que la UVigo es la institución con mayor impacto en la comunidad académica, con un punto de inflexión relevante en el año 2016, cuando alcanzó su mayor número de citas. Le sigue la USC, cuyos años de mayor impacto se registran en 2016 y 2021. Por su parte, la UDC tuvo su año más destacado en términos de citaciones en 2017. En cuanto a los resultados obtenidos mediante OpenAlex, la USC aparece como la institución con mayor volumen de citaciones, con picos importantes en los años 2014 y 2020. En segundo lugar se encuentra la UVigo, con un máximo en 2018. La UDC, en esta fuente, alcanza su punto más alto de impacto en 2022.

Se observa que en todas las BBDD, hay una disminución general del número de citaciones a partir del año 2020. Este descenso podría atribuirse a distintos factores: por ejemplo, podría ser que la repercusión de las publicaciones recientes en la comunidad académica haya sido baja. Pero principalmente es debido al reconocimiento y citaciones, que conllevan algo de tiempo. Es decir, si una publicación X se lanza en un año Y, su impacto dependerá de su relevancia, aunque puede ser que el impacto lleve tiempo para extenderse o consolidarse.

Fragmento de código 3.2: Calcula el número de citas para cada universidad del SUG. Posteriormente, se grafica la evolución de las citas con `ggplot`. Resultados para los repositorios de OpenAlex, Crossref y WoS, respectivamente.

```
# Impacto en términos de citas
impacto_citas_crsrf <- data_sug_crf %>% mutate(Fuente="CrossRef")%>%
     group_by(PY,SUG,Fuente) %>%summarise(total_citas = sum(TC, na.rm = TRUE),
     promedio_citas = mean(TC, na.rm = TRUE),publicaciones = n(),.groups="drop")
impacto_citas_opnlx <- data_sug_opnlx%>%mutate(Fuente = "OpenAlex")%>%
     group_by(PY,SUG,Fuente) %>%summarise(total_citas = sum(TC, na.rm = TRUE),
     promedio_citas = mean(TC, na.rm = TRUE),publicaciones = n(),.groups="drop")
impacto_citas_wos <- data_sug_wos%>%mutate(Fuente = "WoS")%>%
     group_by(PY,SUG,Fuente) %>%summarise(total_citas = sum(TC, na.rm = TRUE),
     promedio_citas = mean(TC, na.rm = TRUE),publicaciones = n(),.groups="drop")

# Ejemplo del gráfico Crossref-Wos
Mee<- bind_rows(impacto_citas_crsrf, impacto_citas_wos)
ggplot(Mee, aes(x = factor(PY), y = log(total_citas),
  color = SUG, linetype = Fuente,group = interaction(SUG, Fuente)))+ geom_line()+
  geom_point(size = 1.5)+labs(x = "Año",y = "log(Total de citas)",color="SUG",
  linetype="Fuente")+theme_minimal()+theme(legend.position="right")

  # Ejemplo del gráfico OpenAlex-Wos
Mee<- bind_rows(impacto_citas_opnlx, impacto_citas_wos)
ggplot(Mee, aes(x = factor(PY), y = log(total_citas),
  color = SUG, linetype = Fuente,group = interaction(SUG, Fuente)))+ geom_line()+
  geom_point(size = 1.5)+labs(x = "Año",y = "log(Total de citas)",color="SUG",
  linetype="Fuente")+theme_minimal()+theme(legend.position="right")
```

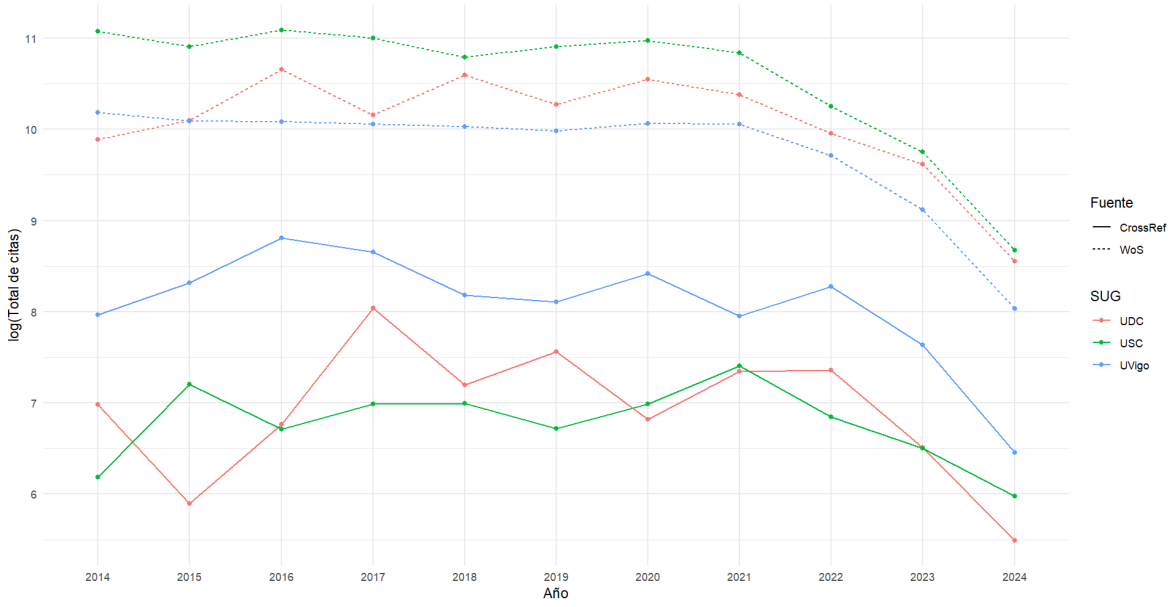

Figura 3.3: Citas recibidas por los documentos, en función del año y universidad, en la base de datos Crossref y WoS.

Tabla 3.2: Citas, índices, y publicaciones, por año y universidad, según la base de datos de OpenAlex

Año (PY)	SUG	OpenAlex				WoS			
		Índice H	Índice G	Total citas	Nº Publicaciones	Índice H	Índice G	Total citas	Nº Publicaciones
2014	UDC	68	106	20458	1023	70	103	19685	780
	UVigo	84	144	39381	1670	70	118	26440	904
	USC	112	194	72258	2523	108	185	64536	1753
2015	UDC	69	109	23614	1308	68	113	24302	972
	UVigo	84	130	36924	1796	70	109	24066	893
	USC	105	173	64460	2754	98	159	54556	1707
2016	UDC	71	105	23966	1329	76	180	42517	1006
	UVigo	80	150	39957	1691	66	107	23957	918
	USC	99	174	61350	2720	97	193	65142	1772
2017	UDC	66	107	23360	1466	65	113	25728	1041
	UVigo	88	137	38957	1835	68	107	23225	938
	USC	95	169	60402	2946	88	186	59807	1866
2018	UDC	69	146	32859	1621	78	173	39835	1086
	UVigo	85	167	45164	1806	63	99	22585	975
	USC	96	152	56783	3243	88	141	48382	1930
2019	UDC	68	127	28200	1819	73	123	28787	1318
	UVigo	80	138	37807	1887	64	90	21585	1070
	USC	99	158	58659	3305	92	162	54460	1976
2020	UDC	65	123	28663	1929	74	158	37999	1387
	UVigo	82	164	50163	2267	58	101	23468	1177
	USC	101	194	73672	3709	91	167	58082	2151
2021	UDC	53	78	16400	1889	68	136	32167	1520
	UVigo	89	127	42311	2550	62	91	23323	1338
	USC	87	163	57193	3749	84	158	50968	2375
2022	UDC	37	57	10205	1609	55	93	21090	1586
	UVigo	68	107	31602	2670	48	78	16520	1297
	USC	58	109	30469	3478	60	101	28242	2175
2023	UDC	26	44	5635	1803	52	85	15007	1515
	UVigo	39	56	13461	2623	34	48	9118	1215
	USC	38	66	15664	3493	43	80	17213	2149
2024	UDC	10	17	1335	1199	26	45	5170	1465
	UVigo	13	18	2421	1880	19	26	3079	1163
	USC	16	32	3954	2795	25	38	5863	2204

Tabla 3.3: Citas, índices, y publicaciones, por año y universidad, según la base de datos de Crossref

Año (PY)	SUG	Crossref				WoS			
		Índice H	Índice G	Total citas	Nº Publicaciones	Índice H	Índice G	Total citas	Nº Publicaciones
2014	UDC	7	32	1067	71	70	103	19685	780
	UVigo	30	49	2861	143	70	118	26440	904
	USC	12	18	484	124	108	185	64536	1753
2015	UDC	8	14	359	138	68	113	24302	972
	UVigo	33	59	4078	222	70	109	24066	893
	USC	16	30	1339	255	98	159	54556	1707
2016	UDC	12	25	854	209	76	180	42517	1006
	UVigo	35	78	6666	244	66	107	23957	918
	USC	14	21	818	252	97	193	65142	1772
2017	UDC	16	51	3051	361	65	113	25728	1041
	UVigo	31	70	5705	339	68	107	23225	938
	USC	15	24	1081	321	88	186	59807	1866
2018	UDC	18	29	1336	431	78	173	39835	1086
	UVigo	31	47	3532	421	63	99	22585	975
	USC	15	24	1077	384	88	141	48382	1930
2019	UDC	17	38	1900	601	73	123	28787	1318
	UVigo	31	45	3289	618	64	90	21585	1070
	USC	14	21	828	395	92	162	54460	1976
2020	UDC	13	21	891	840	74	158	37999	1387
	UVigo	34	54	4498	596	58	101	23468	1177
	USC	17	24	1072	505	91	167	58082	2151
2021	UDC	17	31	1506	624	68	136	32167	1520
	UVigo	27	39	2798	527	62	91	23323	1338
	USC	18	34	1623	587	84	158	50968	2375
2022	UDC	16	31	1459	571	55	93	21090	1586
	UVigo	27	51	3880	624	48	78	16520	1297
	USC	13	18	911	685	60	101	28242	2175
2023	UDC	11	17	646	647	52	85	15007	1515
	UVigo	20	28	2006	687	34	48	9118	1215
	USC	11	16	640	523	43	80	17213	2149
2024	UDC	6	8	222	691	26	45	5170	1465
	UVigo	10	16	599	562	19	26	3079	1163
	USC	9	10	363	1333	25	38	5863	2204

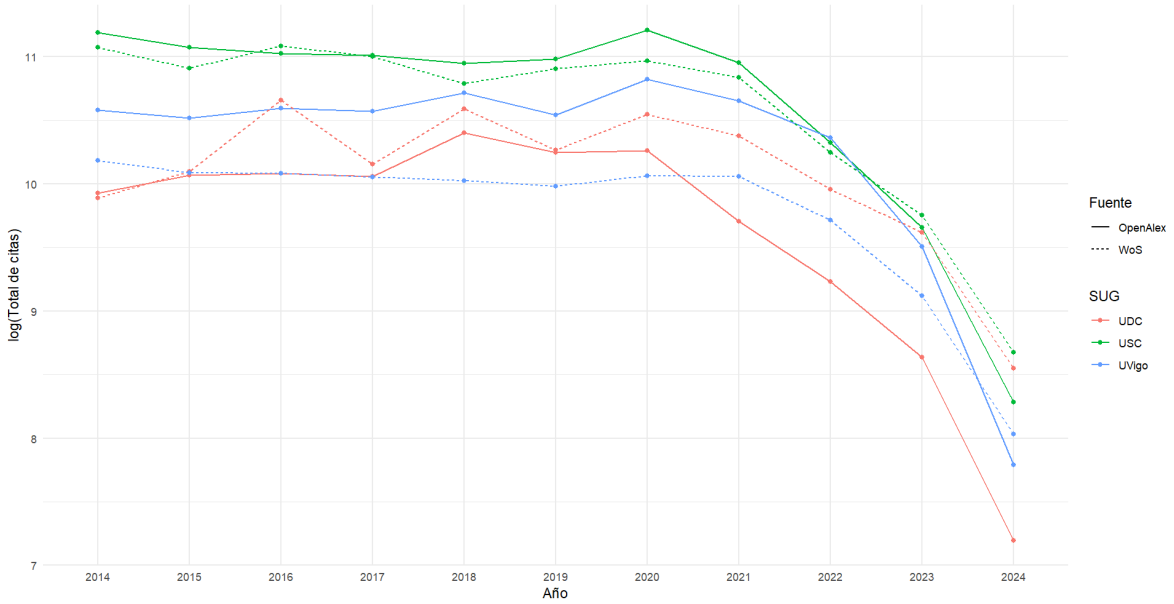

Figura 3.4: Citas recibidas por los documentos, en función del año y universidad, en la base de datos OpenAlex y WoS.

3.2.2. Índices bibliométricos fuente-universidad

Además del número de citas, para completar el análisis del impacto de la producción científica de investigadores e instituciones, es habitual incluir indicadores más recientes como es el caso del índice h y el índice g. Ambos proporcionan información acerca del balance entre citas y número de documentos producidos. En el caso del índice g, se puede considerar como una modificación del índice h que pondera positivamente la existencia de documentos que reciben un número comparativamente elevado de citas, por lo que es mayor, en magnitud, que el índice h. Para su cálculo se ha utilizado el paquete `bibliometrix`, en concreto la función `Hindex()`.

Índice h y g

La Figura 3.6 y la Figura 3.7, muestran, de una forma intuitiva y directa, la información contenida en la Tabla 3.3 y Tabla 3.2, respectivamente. A partir de estos resultados, se observa que en Crossref, la UVigo es la institución que ha alcanzado un mayor impacto académico a lo largo del periodo analizado. Esto se evidencia por el comportamiento ascendente de los índices h y g, especialmente en los años 2014-2017, durante los cuales las citaciones y publicaciones reflejan un crecimiento notable. El descenso posterior en ambos índices podría explicarse por su dependencia directa tanto del número de citaciones totales (TC) como del número de publicaciones (NP). Como se mencionó anteriormente, si una publicación X se genera en el año Y, su impacto estará determinado por su relevancia académica, pero dicho impacto puede tardar en consolidarse a lo largo del tiempo.

En cuanto a los resultados obtenidos a partir de las BBDD de OpenAlex (véase Figura 3.7), la USC se posiciona como la universidad con mayor impacto académico durante el periodo estudiado. Los valores alcanzados en los índices h y g resultan significativamente más altos que los del resto de universidades. Se identifica un punto de inflexión en el año 2020, cuando ambos índices alcanzan su punto máximo, seguido por un ligero descenso. Esta caída, al igual que en el caso de CrossRef, puede deberse a una menor cantidad de citaciones en los años más recientes o a la incorporación de nuevas publicaciones que aún no han tenido tiempo suficiente para ser citadas. A la USC le siguen, en términos de impacto, la UVigo y la UDC, aunque con diferencias más marcadas según el repositorio.

Por otra parte, en los datos provenientes de WoS, los valores de los índices h y g no son tan altos como los calculados a partir de OpenAlex. Se observan unas variaciones o picos, especialmente en los primeros años,

probablemente debido a los criterios más estrictos de indexación que aplica WoS. La USC también lidera en esta base, seguida por la UDC y la UVigo, reflejando un patrón similar al observado en OpenAlex, aunque con mayor variabilidad interanual. En conjunto, los resultados reflejan patrones de evolución diferenciados entre las universidades, así como entre las BBDD lo que pone de manifiesto la importancia de considerar múltiples fuentes para el análisis bibliométrico.

Fragmento de código 3.3: Calcula los índices h y g para cada universidad del SUG, mediante la función `Hindex`. Posteriormente, grafica la evolución de los índices con `ggplot`. Resultados para los repositorios de OpenAlex, Crossref y WoS, respectivamente.

```
# Índice h, g
indices_crf <- do.call(rbind,lapply(min(data_sug_crf$PY):max(data_sug_crf$PY),
    function(x) Hindex(data_sug_crf %>% filter(PY==x) %>%mutate(AU=SUG),field="author",
      years=Inf,sep = ";")$H))
indices_opnlx <- do.call(rbind,lapply(min(data_sug_opnlx$PY):max(data_sug_opnlx$PY),
    function(x) Hindex(data_sug_opnlx %>% filter(PY==x) %>%mutate(AU=SUG),
     field="author",years=Inf,sep = ";")$H))
indices_wos <- do.call(rbind,lapply(min(data_sug_wos$PY):max(data_sug_wos$PY),
    function(x) Hindex(data_sug_wos%>%filter(PY==x) %>%mutate(AU=SUG),
          field="author",years=Inf,sep=";")$H))

# Ejemplo del gráfico Wos
ggplot(indices_wos%>%pivot_longer(cols = c(h_index, g_index),names_to = "Indice",
 values_to = "Valor_Indice"), aes(x = factor(PY_start),y=Valor_Indice, col=Element,
 linetype=Indice,group=interaction(Element,Indice)))+geom_line()+geom_point()+
 labs(x="Año", y="Índice", color="SUG", linetype="Tipo Índice")+theme_minimal()

 # Ejemplo del gráfico OpenAlex
ggplot(indices_opnlx%>%pivot_longer(cols = c(h_index, g_index),names_to = "Indice",
 values_to = "Valor_Indice"), aes(x = factor(PY_start),y=Valor_Indice, col=Element,
 linetype=Indice,group=interaction(Element,Indice)))+geom_line()+geom_point()+
 labs(x="Año", y="Índice", color="SUG", linetype="Tipo Índice")+theme_minimal()
```

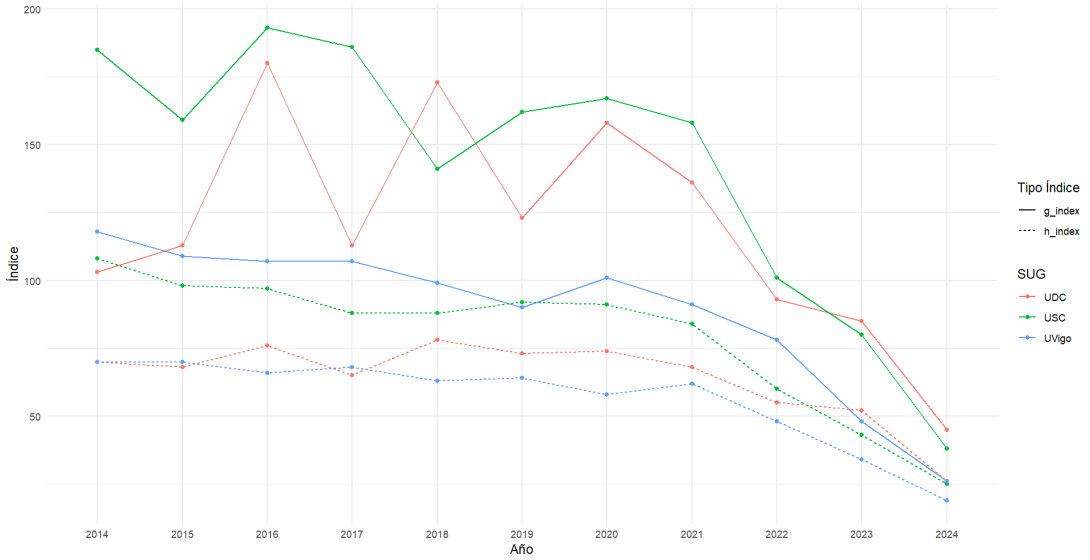

Figura 3.5: Índices h y g obtenidos a partir de la información aportada por WoS.

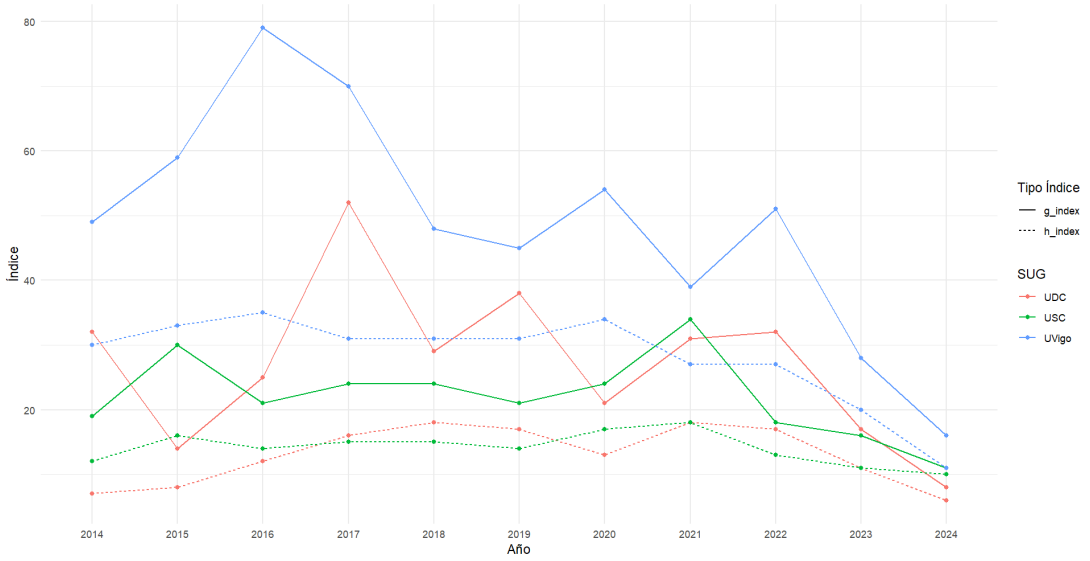

Figura 3.6: Índices h y g obtenidos a partir de la información aportada por CrossRef

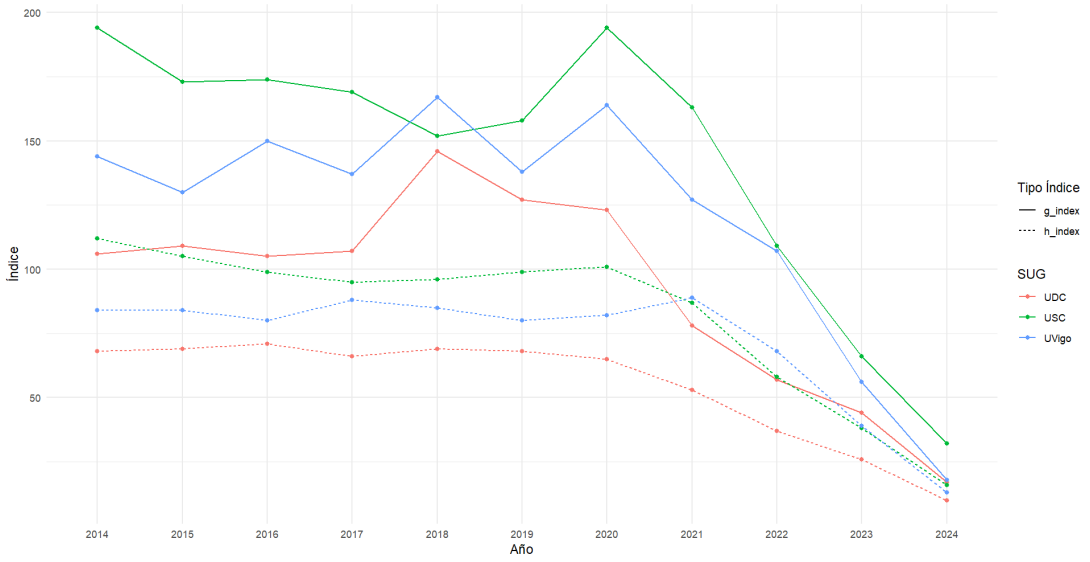

Figura 3.7: Índices h y g obtenidos a partir de la información aportada por OpenAlex

3.3. Patrones de colaboración institucional y de autoría

Un aspecto fundamental en el análisis bibliométrico es la identificación descripción de las redes de investigación, ya sea a nivel de instituciones, trabajos citados o autores, entre otros. En esta sección se identifican las redes de colaboración institucional de cada una de las universidades gallegas, mediante la aplicación de la función `biblioNetwork()`, del paquete `bibliometrix`. Con el objeto de obtener un conjunto de resultados representativo, interpretable e ilustrativo, se han tenido en cuenta aquellas 10 intituciones más relevanes para cada universidad del SUG.

Los resultados obtenidos para la UDC se presentan en la Tabla 3.4 (OpenAlex) y en la Tabla 3.5 (WoS), mientras que la Figura 3.8 se corresponde con la salida gráfica del análisis. Ambas fuentes muestran coincidencias importantes respecto a las instituciones con las que la UDC mantiene vínculos de colaboración, aunque existen diferencias notables en volumen y cobertura que merecerían otro análisis. En ambos casos, se identifica como principal patrón de coautoría el trabajo con investigadores de la propia UDC, lo cual era esperable y habitual. Además, destacan como colaboradores recurrentes: la Universidade de Santiago de Compostela (USC), la Universidade de Vigo (UVigo), el Instituto de Investigación Biomédica de A Coruña, etc. A este respecto, es importante destacar (Figura 3.8) que la UDC destaca sobre todo en colaboración nacional y, es más, de tipo clínico o biomédico, teniendo en cuenta el alto número de hospitales con y centros de investigación de la rama sanitaria con los que colabora. En el caso de la información proporcionada por WoS (Tabla 3.5), hay que indicar que se incluyen centros de investigación de la propia UDC que, por tanto, al no ser entidades aparte, no se deberían considerar; tales son los casos de los centros CITIC y CICA

Estos resultados refuerzan la idea de una red de colaboración sólida dentro del Sistema Universitario de Galicia (SUG), donde las tres universidades gallegas comparten múltiples publicaciones y proyectos conjuntos. El análisis también permite identificar algunas diferencias relevantes entre las BBDD. Por ejemplo, OpenAlex reporta una mayor frecuencia absoluta de colaboraciones (15576 registros con la propia UDC frente a los 4662 de WoS), así como una lista más extensa de instituciones. Esto puede deberse a distintos criterios de indexación, a una mayor amplitud de fuentes cubiertas, o a una mejor integración de datos procedentes de revistas menos tradicionales o de acceso abierto. Así, en OpenAlex aparecen colaboraciones que no se reflejan en WoS, como la Universidad Rey Juan Carlos, Universidad Europea de Madrid, o incluso con hospitales y centros fuera del entorno gallego. Este comportamiento sugiere que OpenAlex podría estar capturando una visión más diversa y reciente de la actividad investigadora de la UDC. Por otro lado, WoS parece ofrecer un panorama más consolidado y probablemente más riguroso en cuanto a estándares de indexación, pero más restringida en volumen. Por ejemplo, en WoS se destaca el Complejo Hospitalario Universitario A Coruña (CHUAC), con una frecuencia de colaboración notable,

que también está presente en OpenAlex, aunque con valores distintos. Esto podría reflejar diferencias en cómo cada base agrupa o identifica a las instituciones colaboradoras.

A pesar de estas variaciones, se observa en general una limitada proyección internacional directa desde la UDC. Si bien algunas instituciones extranjeras como la Universidade do Porto o la University of Barcelona figuran en ambas listas, la mayoría de las colaboraciones se mantienen a nivel España. Esto no quiere decir que la UDC carezca de impacto global, sino que su red de colaboración institucional, al menos la que está registrada en las publicaciones indexadas, está más enfocada en el ámbito nacional y regional. En resumen, tanto WoS como OpenAlex permiten concluir que la UDC mantiene una actividad colaborativa significativa dentro del SUG y con otras instituciones españolas, especialmente en áreas biomédicas. Las diferencias entre ambas fuentes enriquecen el análisis, mostrando que el alcance y la visibilidad de las colaboraciones pueden depender en parte de la base consultada, sus criterios de indexación y la actualidad de sus registros.

Fragmento de código 3.4: Crea una red de colaboración institucional para las universidades del SUG mediante `biblioNetwork`. La función `mcb` (función propia) devuelve una tabla de pesos o frecuencias de las colaboraciones, solo se utiliza con OpenAlex. WoS al tener un formato estándar, funciona perfectamente con `bibliometrix`. Despues se grafican las colaboraciones mediante `networkPlot`

```
# Red de colaboración institucional: OpenAlex
NetMatrix_Udc <- biblioNetwork(Udc_M_Opnlx,analysis="collaboration",
                    network="universities", sep=";")
# Función de colaboración institucional: Devulve un matriz de pesos(Frecuencias)
mcb <- function(NetMatrix,ref="UNIVERSIDADE DE VIGO"){
     cdf <- as.data.frame(as.matrix(NetMatrix)); cdf$Institution <- rownames(cdf)
     cruz <- pivot_longer(cdf,cols=-Institution,names_to="Collaborator",
     values_to="Frequency")%>% filter(Institution==ref)
     # Sum frecuencias para el total de colaboraciones por institución
      colab_total <- cruz %>% group_by(Institution, Collaborator) %>%
        summarise(Total_Frequency = sum(Frequency, na.rm = TRUE)) %>%
        ungroup()%>%arrange(desc(Total_Frequency))
      return(colab_total)}
a <- mcb(NetMatrix_Udc,ref="UNIVERSIDADE DA CORUÑA") # Tabla OpenAlex
resA <- biblioAnalysis(udc_M_wos);head(resA$Aff_frac,10) # Tabla Wos

# Ejemplo gráfica colaboraciones UDC: OpenAlex
g <- graph_from_data_frame(a[1:15,],directed = FALSE)
E(g)$weight <- a[1:15,]$Total_Frequency; V(g)$size <- strength(g,mode="all")*0.5
ggraph(as_tbl_graph(g),layout="fr")+geom_edge_link(aes(width=E(g)$weight/100),
            color="gray", alpha=0.7) + # Aristas con grosor según frecuencia
geom_node_point(aes(size = V(g)$size), color = "lightblue") + # Tamaño del nodo
geom_node_text(aes(label = name), repel = TRUE, size = 3, color = "black") +#Etiquets
  theme_minimal() + labs() +theme(legend.position = "none")
```

Tabla 3.4: Principales instituciones colaboradoras con la UDC, OpenAlex.

Institución	Colaborador	Frecuencia
UDC	UNIVERSIDADE DA CORUÑA	15576
	UNIVERSIDADE DE SANTIAGO DE COMPOSTELA	1344
	UNIVERSIDADE DE VIGO	594
	INSTITUTO DE INVESTIGACIÓN BIOMÉDICA DE A CORUÑA	511
	UNIVERSIDAD COMPLUTENSE DE MADRID	293
	COMPLEXO HOSPITALARIO UNIVERSITARIO A CORUÑA	282
	UNIVERSIDAD REY JUAN CARLOS	258
	UNIVERSIDAD DE OVIEDO	257
	HOSPITAL UNIVERSITARIO PUERTA DE HIERRO MAJADAHONDA	177
	UNIVERSIDAD DE LEÓN	172
	UNIVERSIDADE DO PORTO	163
	UNIVERSIDAD DE SALAMANCA	160
	UNIVERSIDAD EUROPEA DE MADRID	155

Tabla 3.5: Principales instituciones colaboradoras con la UDC, WoS.

Institución	Colaborador	Frecuencia
UDC	UNIVERSIDADE DA CORUÑA	4662
	COMPLEJO HOSPITALARIO UNIVERSITARIO A CORUÑA	634
	UNIVERSIDADE DE VIGO	438
	UNIVERSIDADE DE SANTIAGO DE COMPOSTELA	397
	INSTITUTO DE INVESTIGACIÓN BIOMÉDICA DE A CORUÑA (INIBIC)	349
	CENTRO INTERDISCIPLINAR DE QUÍMICA E BIOLOXÍA (CICA)	256
	CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS (CSIC)	129
	UNIVERSITY OF BARCELONA	94
	UNIVERSIDADE DO PORTO	94
	CIBER - CENTRO DE INVESTIGACIÓN BIOMÉDICA EN RED	92

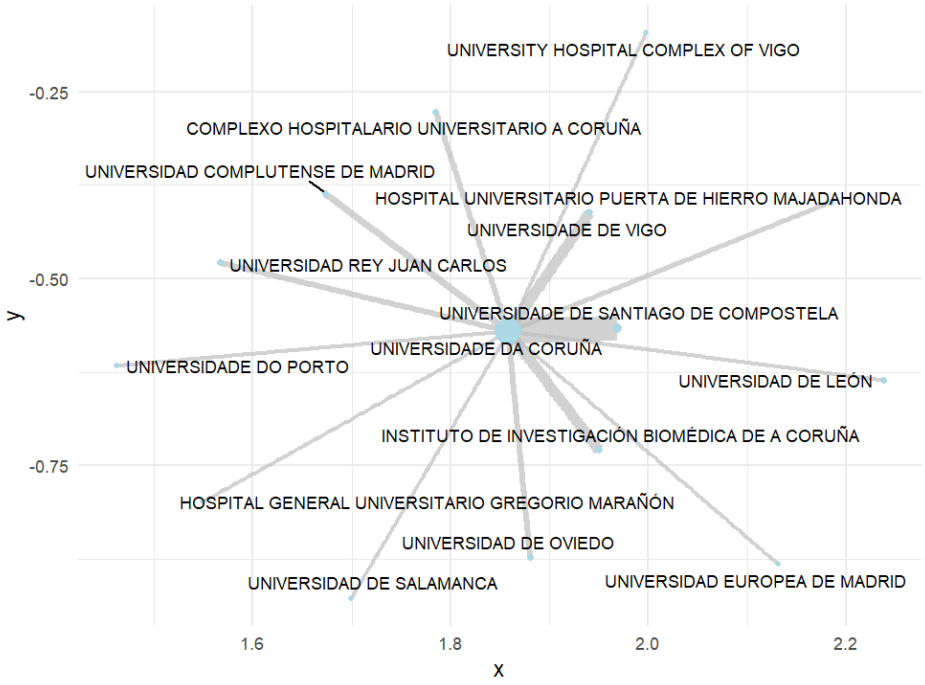

Figura 3.8: Colaboraciones institucionales en la UDC, OpenAlex

Los resultados obtenidos para la USC se presentan en la Tabla 3.6 (OpenAlex) y en la Tabla 3.7 (WoS), mientras que la Figura 3.9 representa la correspondiente salida gráfica. El grosor de cada línea entre dos instituciones es proporcional al número de documentos en común publicados. Además, estarán en una posición más central aquellas instituciones que aparezcan más en los documentos firmados. Las principales líneas de cooperación institucional identificadas son las siguientes: otros autores de la misma USC, autores de la UVigo, autores de la UDC, Investigadores del Instituto de Investigación Sanitaria de Santiago (IDIS), etc. En ambas bases de datos, la USC destaca como un nodo central de colaboración en el entorno académico gallego, evidenciando una fuerte integración con el resto del SUG. Sin embargo, similar al caso anterior, los resultados muestran diferencias que podrían enriquecer y/o expandir la interpretación. En WoS, se observa una concentración significativa en colaboraciones con instituciones locales y nacionales, como el Complejo Hospitalario Universitario de Santiago (CHUS), el CSIC, y centros especializados como el CIQUS (este caso particular no es una institución aparte de la USC, de hecho es un centro de investigación de la misma).

Estas colaboraciones reflejan una relación con un entorno sanitario y biomédico, así como con centros propios de la universidad. Por su parte, OpenAlex muestra una red más extensa. Destacando el alto número de colaboraciones con instituciones internacionales, lo cual resaltaría a la USC. Se observan colaboraciones frecuentes con entidades como el CERN (European Organization for Nuclear Research), la University of Manchester, el CNRS francés y varias universidades europeas (Heidelberg, Liverpool, Barcelona). También se incluyen centros latinoamericanos como la Universidade Federal do Rio de Janeiro. Esto sugiere una internacionalización y participación activa en proyectos globales, en áreas de ciencia básica, medicina y física. Comparando los volúmenes, OpenAlex reporta una frecuencia superior (por ejemplo, más de 30000 colaboraciones con la propia USC), frente a los casi 10000 que reporta WoS. Esta diferencia puede explicarse por la mayor cobertura de OpenAlex, que incluye repositorios, preprints, etc. Asimismo, OpenAlex tiende a reflejar con mayor detalle las estructuras internas de las universidades y centros adscritos, como el CIQUS.

Ambos conjuntos de datos confirman que la USC mantiene una colaboración activa y diversa, tanto a nivel local (SUG), a nivel nacional (España) y a nivel internacional (mayormente Europa). Es en este ámbito en el que se diferencia con respecto a la UDC; es decir, la investigación de la USC parece que está más internacionalizada. Este aspecto resulta positivo en el ámbito de la investigación (la proporción de internacionalización comúnmente está directamente relacionada con el impacto).

Fragmento de código 3.5: Ejemplo de aplicación para la USC. Repositorios OpenAlex y WoS

```
# Red de colaboración institucional: OpenAlex
NetMatrix_Usc <- biblioNetwork(Usc_M_Opnlx,analysis="collaboration",
                    network="universities", sep=";")
b <- mcb(NetMatrix_Udc,ref="UNIVERSIDADE DA CORUÑA") # Tabla OpenAlex
resB <- biblioAnalysis(usc_M_wos);head(resB$Aff_frac,10) # Tabla Wos
```

Tabla 3.6: Principales colaboradores institucionales de la USC, OpenAlex

Institución	Colaborador	Frecuencia
USC	UNIVERSIDADE DE SANTIAGO DE COMPOSTELA	30981
	UNIVERSIDADE DE VIGO	1623
	UNIVERSIDADE DA CORUÑA	1344
	INSTITUTO DE INVESTIGACIÓN SANITARIA DE SANTIAGO	1113
	UNIVERSITAT DE BARCELONA	877
	CENTER FOR RESEARCH IN MOLECULAR MEDICINE AND CHRONIC DISEASES	869
	EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH	865
	UNIVERSITY OF MANCHESTER	775
	CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE	747
	UNIVERSIDADE FEDERAL DO RIO DE JANEIRO	687
	UNIVERSITY OF LIVERPOOL	663
	NATIONAL INSTITUTE FOR SUBATOMIC PHYSICS	648
	HEIDELBERG UNIVERSITY	636

Tabla 3.7: Principales instituciones colaboradoras con la USC, WoS.

Institución	Colaborador	Frecuencia
USC	UNIVERSIDADE DE SANTIAGO DE COMPOSTELA	9794
	COMPLEXO HOSPITALARIO UNIVERSITARIO DE SANTIAGO DE COMPOSTELA	728
	UNIVERSIDADE DE VIGO	447
	UNIVERSIDADE DA CORUÑA	343
	CIBER - CENTRO DE INVESTIGACIÓN BIOMÉDICA EN RED	278
	CIQUS	263
	CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS (CSIC)	253
	UNIVERSIDADE DO PORTO	207
	UNIVERSITY OF BARCELONA	133
	CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE (CNRS)	119

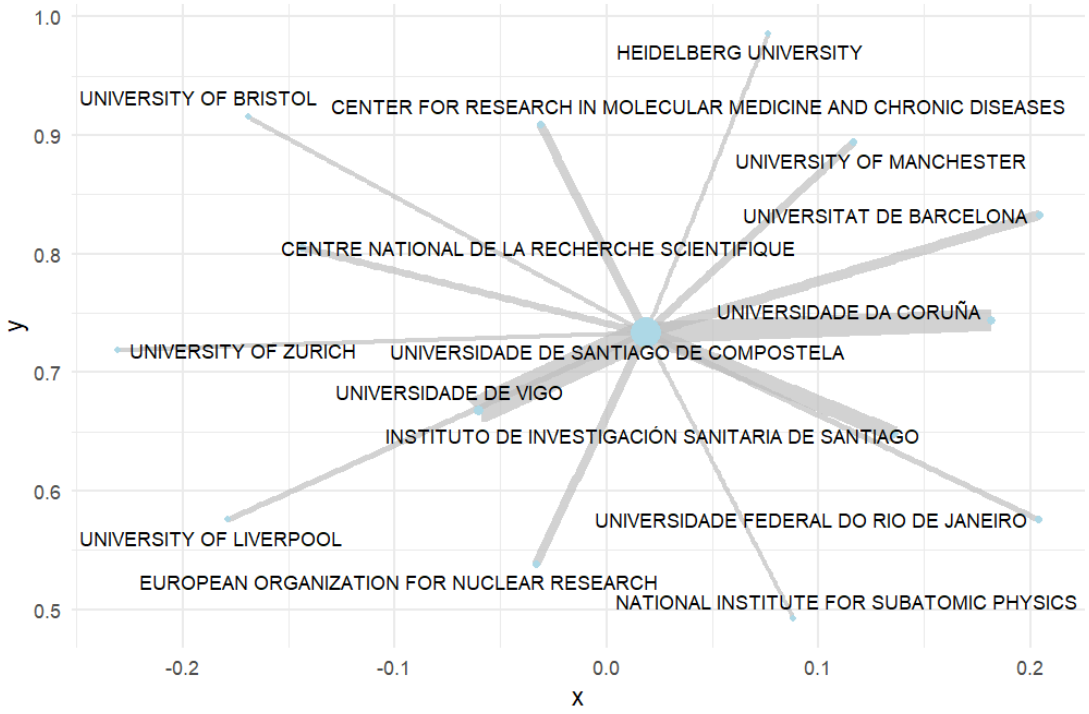

Figura 3.9: Colaboraciones institucionales en la USC, OpenAlex

Para la UVigo, los resultados se exponen en la Tabla 3.9 (OpenAlex) y en la Tabla 3.8 (WoS). Una representación visual de estas colaboraciones se encuentra también en la Figura 3.10. Las principales líneas de colaboración institucional incluyen: Autores de la misma UVigo, autores de la USC, autores de la UDC, investigadores de la Universidade do Porto, etc. Los datos confirman que la UVigo mantiene una red de colaboración consolidada principalmente dentro del SUG, y en menor medida, a nivel nacional e internacional. En ambas BBDD se evidencia que la mayor parte de las colaboraciones se concentra dentro de la propia institución.

En el caso de WoS, el análisis refleja una estructura de colaboración más centrada en el ámbito local y regional. Destacan los centros de investigación propios de la UVigo, como el CINBIO, CIM UVIGO, CINTECX y ATLANTTIC, lo que sugiere una fuerte integración entre facultades y centros de investigación internos. Asimismo, se identifican colaboraciones con otras instituciones gallegas como la USC, la UDC e instituciones extranjeras (como Portugal), la Universidade do Porto y la Universidade do Minho. Esto apunta a un enfoque regional de colaboración, aunque con lazos hacia Portugal, tal vez por razones lingüísticas, culturales y/o de proximidad geográfica. Por otra parte, los datos extraídos de OpenAlex ofrecen una visión más amplia y con mayor volumen de colaboraciones. OpenAlex reporta más de 21000 vínculos con la propia UVigo, lo que sugiere una mayor sensibilidad en el rastreo de coautorías internas y una cobertura más extensa de publicaciones, incluyendo posiblemente trabajos en repositorios y conferencias no indexadas en WoS. Se destacan colaboraciones con universidades extranjeras: University of Aveiro y University of Lisbon, así como instituciones españolas como la Universidad de Granada, la Universidad Complutense de Madrid y la Universidad de Extremadura.

Se identifican vínculos con el Servicio Gallego de Salud y la Galicia Sur Biomedical Foundation, lo que refuerza la idea de colaboración activa en los campos de la biomedicina, la ingeniería y las ciencias marinas. La presencia del Instituto de Investigacións Mariñas sugiere colaboraciones en áreas marítimas y medioambientales, coherente con el contexto geográfico de la institución. Aunque el volumen de colaboraciones internacionales es inferior al nacional o regional, no quiere decir que no haya colaboraciones internacionales; las colaboraciones internacionales no son tan frecuentes o no se han registrado. Aun así, la UVigo muestra una proyección de expansión, especialmente hacia Portugal. Se puede observar que en la UVigo, al igual que en la USC, la investigación está más internacionalizada que en la UDC, en la que prima la investigación conjunta con centros españoles del ámbito biosanitario y clínico.

Fragmento de código 3.6: Ejemplo de aplicación para la Uvigo. Repositorios OpenAlex y WoS

```
# Red de colaboración institucional: OpenAlex
NetMatrix_Udv <- biblioNetwork(Udv_M_Opnlx,analysis="collaboration",
                    network="universities", sep=";")
c1 <- mcb(NetMatrix_Udv,ref="UNIVERSIDADE DE VIGO") # Tabla OpenAlex
resC <- biblioAnalysis(udv_M_wos);head(resC$Aff_frac,10) # Tabla Wos
```

Tabla 3.8: Principales instituciones colaboradoras con la UVigo, WoS.

Institución	Colaborador	Frecuencia
Uvigo	UNIVERSIDADE DE VIGO	6513
	UNIVERSIDADE DE SANTIAGO DE COMPOSTELA	420
	UNIVERSIDADE DO PORTO	164
	CINBIO	155
	CIM UVIGO	147
	CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS (CSIC)	145
	CINTECX	128
	UNIVERSIDADE DA CORUÑA	121
	ATLANTTIC	98
	UNIVERSIDADE DO MINHO	94

Tabla 3.9: Principales colaboradores institucionales de la UVigo, OpenAlex

Institución	Colaborador	Frecuencia
Uvigo	UNIVERSIDADE DE VIGO	21091
	UNIVERSIDADE DE SANTIAGO DE COMPOSTELA	1623
	UNIVERSIDADE DA CORUÑA	594
	UNIVERSIDADE DO PORTO	416
	UNIVERSITY OF MINHO	342
	UNIVERSIDAD DE EXTREMADURA	269
	UNIVERSITY OF AVEIRO	262
	UNIVERSIDAD DE GRANADA	259
	UNIVERSIDAD COMPLUTENSE DE MADRID	237
	UNIVERSITY OF LISBON	237
	SERVICIO GALLEGO DE SALUD	225
	INSTITUTO DE INVESTIGACIÓNS MARIÑAS	218
	GALICIA SUR BIOMEDICAL FOUNDATION	214

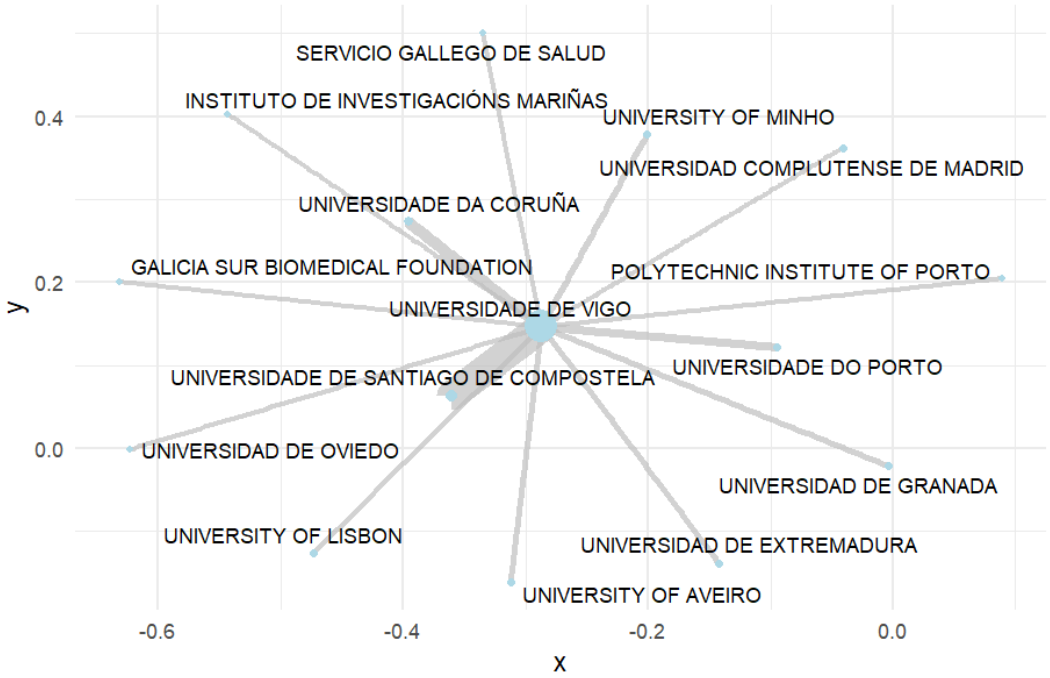

Figura 3.10: Colaboraciones institucionales en la UVigo, OpenAlex

3.4. Áreas temáticas de relevancia y evolución de contenido

Para identificar las principales áreas de investigación en las universidades del SUG, se utilizarán técnicas de co-ocurrencia de palabras clave en su producción científica mediante la función `biblioNetwork()`. El resultado se muestra según un gráfico en el que los nodos representan palabras clave y los enlaces indican la frecuencia con la que estos términos aparecen juntos en los documentos. Además, el tamaño de letra está relacionado con las veces en la que el tema, disciplina o palabra clave aparece en la base de datos estudiada.

Áreas de investigación en la UDC

Las principales áreas de investigación se pueden observar en la Figura 3.11, que muestra una red temática donde los nodos representan disciplinas y los enlaces indican relaciones de co-ocurrencia en publicaciones. El tamaño del nodo refleja la relevancia o frecuencia de aparición de la disciplina, mientras que el grosor del enlace representaría la intensidad de la relación entre ellas.

Se identifican distintos clústeres temáticos:

- **Ciencias de la computación y tecnología**: Esta es el área predominante dentro de la UDC, agrupando disciplinas como *computer science*, *artificial intelligence*, *engineering*, *mathematics* y *physics*. Estas conexiones indican una sólida producción científica en tecnologías digitales, inteligencia artificial y modelado matemático aplicado.
- **Ciencias sociales y humanidades**: Se identifica un clúster verde con conexiones entre *political science*, *business*, *geography* y *sociology*, lo cual sugiere que también existe una línea activa de investigación en ciencias sociales, aunque con menor intensidad comparado con ciencias técnicas.
- **Ciencias de la salud**: Aunque no forma un clúster tan visual como los anteriores, disciplinas como *medicine*, *internal medicine* y *psychology* están conectadas y tienen un papel destacado en el mapa, reflejando su relevancia dentro de la producción científica institucional.

También se observan otras áreas representadas, pero con menor centralidad, como *biology* y *chemistry*, que aparecen vinculadas pero no integradas en un cluster o agrupación. Esto sugeriría que estas disciplinas están presentes, pero de forma más dispersa o con menor colaboración interna.

Fragmento de código 3.7: Crea una red o grafo de asociación de las áreas temáticas de relevancia para las universidades del SUG mediante `biblioNetwork`. Posteriormente, grafica esa red mediante `networkPlot`. Resultados para OpenAlex.

```
# Nota: La generación connlleva algo de tiempo (aprox>12min)
Netrix_Udc_opnlx <- biblioNetwork(Udc_M_Opnlx,analysis="co-occurrences",
                            network="keywords", sep=";")
# Ejemplo Gráfico UDC
# Nota: La generación connlleva algo de tiempo (aprox>12min)
Net_Udc_Coword <- networkPlot(Netrix_Udc_opnlx,normalize="association",n=15,
  type = "fruchterman", size.cex=TRUE, remove.multiple=F,edgesize =4,
  labelsize=2.5,label.cex=TRUE,label.n=15,Title="")
```

Áreas de investigación en la USC

El análisis de co-ocurrencia de palabras clave en la producción científica de la USC se puede observar en la Figura 3.12. Se identifican dos grandes agrupaciones o áreas temáticas (tipo clúster):

- **Ciencias de la vida y la salud**: Este clúster (grupo rojo) incluye disciplinas con mayor peso como *medicine* y *biology*, además de otras áreas importantes como *psychology*, *environmental science* e *internal medicine*. También se encuentran conectadas disciplinas de ciencias exactas como *mathematics*, *physics* y *computer science*. Esto sugiere un enfoque interdisciplinario en investigación biomédica, salud pública y medio ambiente, respaldado por fundamentos matemáticos y computacionales.
- **Ciencias químicas y de materiales**: Este segundo clúster (grupo azul) agrupa disciplinas como *chemistry*, *biochemistry*, *organic chemistry* y *materials science*, lo que indica una especialización en química aplicada, procesos y desarrollo de nuevos materiales con posibles aplicaciones en áreas industriales, médicas o tecnológicas.

Áreas de investigación en la UVigo

Las principales áreas de investigación se pueden observar en la Figura 3.13, que muestra una red temática donde los nodos representan disciplinas y los enlaces indican la frecuencia con la que estos términos aparecen juntos en los documentos o publicaciones.

Se identifican dos grandes agrupaciones o áreas temáticas (tipo clúster):

- **Ciencias exactas y aplicadas**: Representadas en el clúster azul, se observan campos disciplinarios como *computer science*, *engineering*, *mathematics*, *physics*, *environmental science* y *business*, lo que sugiere un fuerte enfoque en tecnología, ciencias aplicadas y sostenibilidad.
- **Ciencias de la vida y la salud**: Representadas en el clúster rojo, incluye áreas como *biology*, *chemistry*, *medicine*, *biochemistry* y *materials science*, lo cual indica una especialización en ciencias de la vida y química aplicada con posibles conexiones hacia biomedicina o biotecnología.

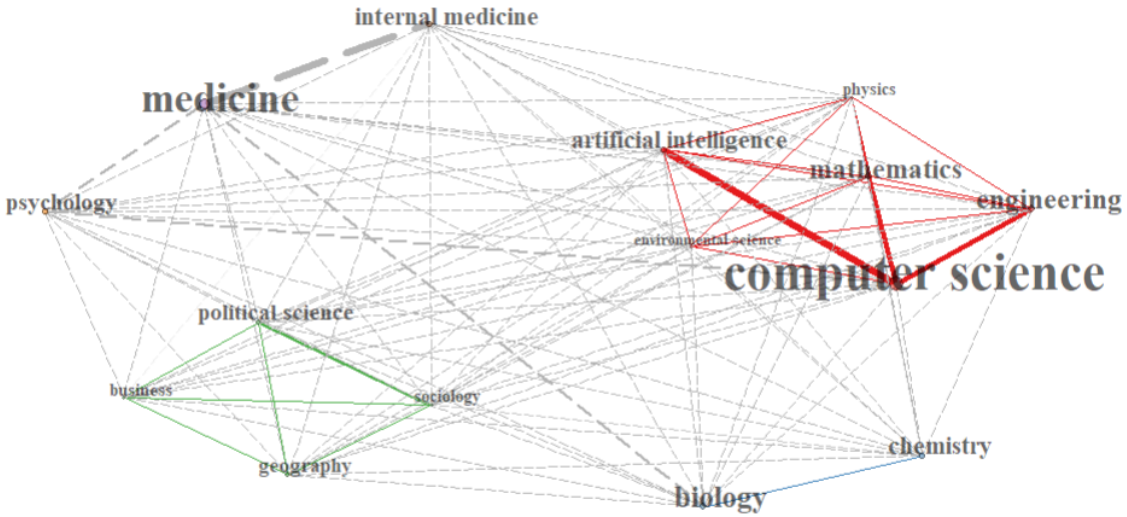

Figura 3.11: Áreas temáticas UDC, OpenAlex

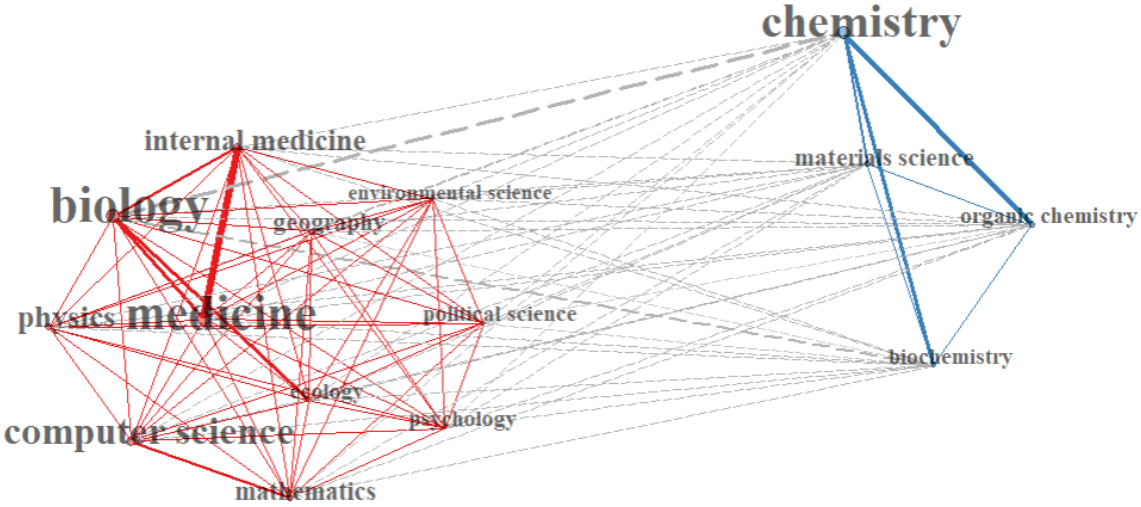

Figura 3.12: Áreas temáticas USC, OpenAlex

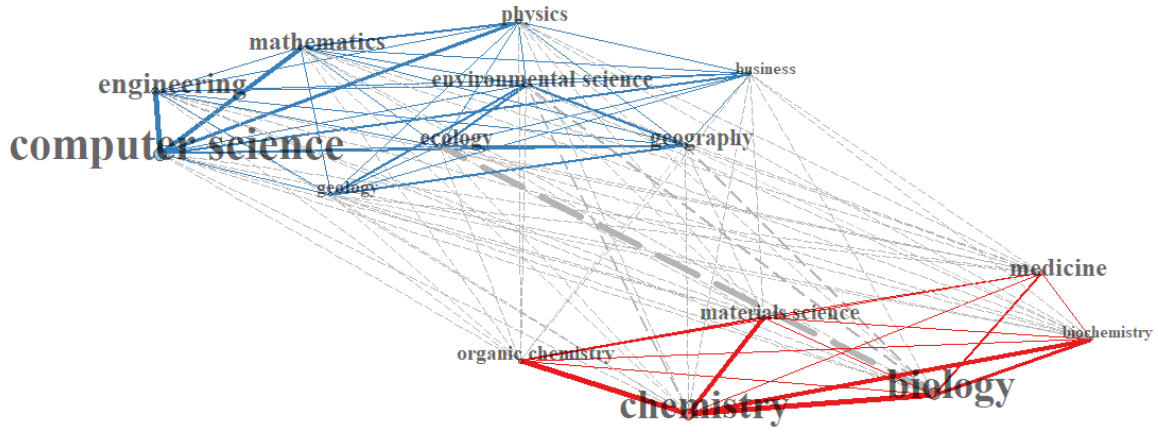

Figura 3.13: Áreas temáticas UVigo, OpenAlex

Una perspectiva general de los campos de investigación en el SUG se presenta en la Tabla 3.10, donde se resumen las áreas principales y subáreas más relevantes identificadas a partir del análisis de co-ocurrencia de palabras clave en OpenAlex.

Tabla 3.10: Comparación de áreas temáticas de investigación en la UVigo, USC y UDC según co-ocurrencia de palabras clave (OpenAlex), obtenidas de las figuras: 3.11, 3.12, y 3.13.

SUG	Áreas principales	Subáreas relevantes
UVigo	Ciencias Computacionales, Ingenierías, Ciencias Naturales	Informática, Matemáticas, Biología, Química, Ciencias Ambientales
USC	Biomedicina, Química/Materiales, Ciencias Sociales	Medicina Interna, Bioquímica, Ciencias Políticas, Psicología, Geografía
UDC	Tecnologías Digitales, Ciencias Sociales, Ingeniería Ambiental	Inteligencia Artificial, Ciencias de la Computación, Negocios, Sociología, Ciencias Ambientales

3.5. Producción del tipo de documento

En esta sección, se pretende estudiar si existe alguna relación estadísticamente significativa entre el tipo de documento que se publica respecto a la universidad de origen (SUG). Interesará contrastar la siguiente hipótesis:

$$\begin{cases} H_0 : \text{No existe asociación entre la universidad y el tipo de documento; } (\textit{Independencia}). \\ H_1 : \text{Existe una asociación entre la universidad y el tipo de documento; } (\textit{Dependencia}). \end{cases} \tag{3.1}$$

Por tanto, para probar H_0 (3.1), se necesitará aplicar un contraste de independencia. Posteriormente, con el análisis de correspondencias (AC), se visualizará, en caso de que exista dependencia, qué tipo de documentos se relacionan con qué universidades.

Fragmento de código 3.8: Realiza una tabla de frecuencias de las categorías filas y columnas, permitiendo ver la distribución de tipo de documento por universidad de estudio. A su vez, permitiría ver si existen frecuencias muy bajas que interferirían con el AC, respectivamente para OpenAlex y WoS.

```
tabla_opnlx <- table(data_sug_opnlx$SUG,data_sug_opnlx$DT_aux)
    kable(tabla_opnlx,format = "latex", booktabs = TRUE)

tabla_wos <- table(data_sug_wos$SUG, data_sug_wos$DT_aux)
    kable(tabla_wos,format = "latex", booktabs = TRUE)
```

Tabla 3.11: Producción del tipo de documento por universidad, OpenAlex

SUG	Artículo o revisión	Difusión sin revisión	Libro o Capítulo	No evaluable	Otro
UDC	14423	754	1381	375	62
USC	30622	1333	1655	1027	78
UVigo	19869	1028	1328	345	105

Tabla 3.12: Producción del tipo de documento por universidad, WoS

SUG	Artículo o revisión	Difusión sin revisión	Libro o Capítulo	No evaluable	Otro
UDC	11424	1483	49	704	16
USC	19840	988	172	1025	33
UVigo	11272	199	81	326	10

Fragmento de código 3.9: Realiza un test de independencia para las tablas de contingencia anteriores, respectivamente para OpenAlex y WoS.

```
# Prueba de hipotesis
chisq.test(tabla_opnlx)
```

```
    Pearson's Chi-squared test

data:  tabla_opnlx
X-squared = 400.46, df = 8,
p-value < 2.2e-16
```

```
# Prueba de hipotesis
chisq.test(tabla_wos)
```

```
    Pearson's Chi-squared test

data:  tabla_wos
X-squared = 1262.2, df = 8,
p-value < 2.2e-16
```

Tomando como referencia el `p-value` de las salidas anteriores, y con un nivel de confianza del 95 %, no se acepta H_0. Es decir, existe una relación significativa entre las universidades y el tipo de documento (H_1). Como existe una relación, se procederá con el análisis de correspondencias para visualizar aquellas categorías que tienen relación y/o aportan a la dependencia (inercias).

Fragmento de código 3.10: Obtiene un resumen o descripciones generales del análisis de correspondencias, como la varianza acumulada, coordenadas de las componentes principales, porcentaje de inercia, etc., respectivamente para los repositorios de OpenAlex y WoS.

```
# AC OpenAlex
res.ac.opnlx<- CA(tabla_opnlx, graph = F); summary(res.ac.opnlx)
```

```
Eigenvalues
                       Dim.1   Dim.2
Variance               0.004   0.001
% of var.             75.754  24.246
Cumulative % of var.  75.754 100.000

Rows
                        Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr   cos2
UDC                   |     2.201 |  0.088 43.045  0.798 |  0.044 34.107  0.202 |
USC                   |     2.038 | -0.065 48.388  0.968 |  0.012  4.943  0.032 |
UVigo                 |     1.145 |  0.034  8.567  0.305 | -0.051 60.949  0.695 |

Columns
                        Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr   cos2
Artículo o revisión   |     0.203 | -0.013  3.671  0.738 | -0.008  4.076  0.262 |
Difusión sin revisión |     0.256 |  0.071  5.244  0.836 | -0.032  3.210  0.164 |
Libro o Capítulo      |     2.949 |  0.210 63.157  0.873 |  0.080 28.604  0.127 |
No evaluable          |     1.646 | -0.199 22.865  0.566 |  0.174 54.692  0.434 |
Otro                  |     0.329 |  0.250  5.063  0.627 | -0.193  9.418  0.373 |
```

```
# Ac Wos
res.ac.wos <- CA(tabla_wos, graph = F); summary(res.ac.wos)
```

```
Eigenvalues
                       Dim.1   Dim.2
Variance               0.026   0.001
% of var.             96.536   3.464
Cumulative % of var.  96.536 100.000

Rows
                        Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr   cos2
UDC                   |    15.932 |  0.235 61.936  0.995 | -0.017  9.347  0.005 |
USC                   |     1.368 | -0.044  3.549  0.664 |  0.032 50.132  0.336 |
UVigo                 |     9.203 | -0.188 34.516  0.960 | -0.039 40.521  0.040 |

Columns
                        Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr   cos2
Artículo o revisión   |     1.942 | -0.046  7.480  0.986 | -0.006  3.044  0.014 |
Difusión sin revisión |    21.985 |  0.626 85.770  0.998 | -0.027  4.406  0.002 |
Libro o Capítulo      |     0.508 | -0.235  1.366  0.688 |  0.158 17.297  0.312 |
No evaluable          |     2.011 |  0.179  5.374  0.684 |  0.121 69.231  0.316 |
Otro                  |     0.058 |  0.044  0.009  0.042 |  0.211  6.022  0.958 |
```

El % de varianza explicada por las componentes (dimensiones), se describe a continuación:

- En OpenAlex, la dimensión 1 explica $\approx 75{,}75\,\%$ de la variabilidad, la dimensión 2 $\approx 24{,}25\,\%$

- En WoS, la dimensión 1 explica $\approx 96{,}5\,\%$, la dimensión 2 $\approx 3{,}5\,\%$.

Tanto en OpenAlex como en WoS, la mayoría de la asociación se captura en la primera dimensión; es más notable en WoS.

Fragmento de código 3.11: Grafica la asociación de las categorías filas y columnas para los repositorios de OpenAlex y WoS, respectivamente.

```
# Representación AC OpenAlex
fviz_ca_biplot(res.ac.opnlx, repel = F, col.row = "steelblue", col.col = "darkred")
# Representación AC WoS
fviz_ca_biplot(res.ac.wos, repel = F,col.row = "steelblue", col.col = "darkred")
```

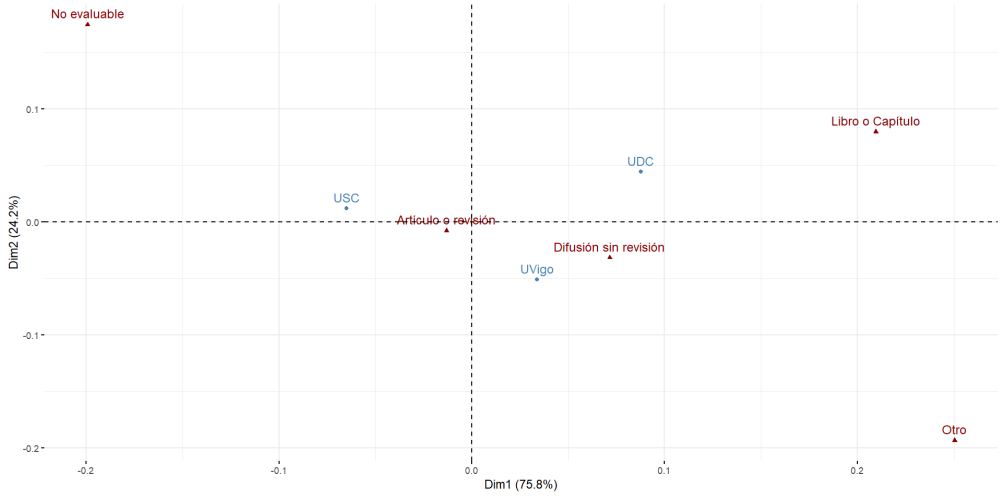

Figura 3.14: Representación de la asociación de las categorías filas y columnas, OpenAlex

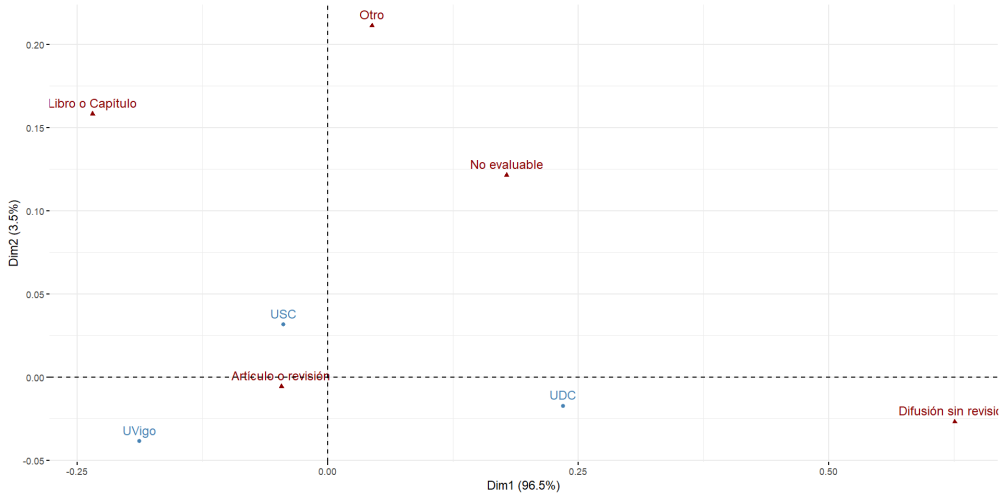

Figura 3.15: Representación de la asociación de las categorías filas y columnas, WoS

Por resolución o compresión de las figuras 3.14 y 3.15, las cercanías no se aprecian muy bien. Para ver de mejor manera las categorías que están más cercanas, se mostrará la distancia euclídea entre los puntos:

Fragmento de código 3.12: Calcula la distancia euclídea entre las categorías filas y columnas del repositorio de OpenAlex

```
# Distancia entre todos los puntos, OpenAlex
coords_todos <- rbind(res.ac.opnlx$row$coord,res.ac.opnlx$col$coord)
dist_todos <- dist(coords_todos)
as.matrix(round(dist_todos,3))[1:3,-c(1:3)]
```

```
    Artículo o revisión Difusión sin revisión Libro o Capítulo No evaluable  Otro
UDC               0.113                 0.077            0.127        0.315 0.288
USC               0.055                 0.143            0.283        0.211 0.376
UVigo             0.064                 0.042            0.219        0.324 0.259
```

En OpenAlex, se tendrían los siguientes resultados:

- La UVigo y la USC están más cerca de la categoría «Artículo o revisión», mientras que la UDC, es más cercana a «Difusión sin revisión».
- Respecto al resto de universidades, la UDC es relativamente más cercana a la categoría «Libro o Capítulo». Esto puede ser debido a la importancia de la ingeniería informática en el ámbito de la UDC, siendo habitual la publicación de conference papers, capítulos de libro relacionados con congresos, etc.
- Las categorías «No evaluable» y «Otro», al estar separadas de los perfiles filas, indicarían que no hay una relación clara de esas categorías con las filas (universidades).

Fragmento de código 3.13: Calcula la distancia euclídea entre las categorías filas y columnas del repositorio de WoS

```
# Distancia entre todos los puntos, WoS
coords_todos <- rbind(res.ac.wos$row$coord,res.ac.wos$col$coord)
dist_todos <- dist(coords_todos)
as.matrix(round(dist_todos,3))[1:3,-c(1:3)]
```

	Artículo o revisión	Difusión sin revisión	Libro o Capítulo	No evaluable	Otro
UDC	0.281	0.391	0.501	0.15	0.298
USC	0.037	0.672	0.229	0.24	0.200
UVigo	0.146	0.814	0.202	0.40	0.341

En WoS, se tendrían los siguientes resultados:

- La USC está más cerca de la categoría «Artículo o revisión», le sigue la UVigo con una distancia algo más grande. Y con mayor distancia la UDC.
- Las categorías «No evaluable», «Otro» y «Libro o Capítulo», al estar separadas de los perfiles filas, indicarían que no hay una relación clara de esas categorías con las universidades. La inercia para «difusión sin revisión» es de $\approx 21{,}98$, lo que indicaría que esta aumenta la varianza en el sistema, explicando así su lejanía respecto al origen.

Capítulo 4

Conclusiones y líneas futuras de investigación

En este trabajo de investigación se han cumplido los objetivos propuestos relacionados con la recolección, procesamiento y análisis exploratorio de datos bibliométricos a través de fuentes abiertas y bases indexadas. A partir del desarrollo y resultados obtenidos, se pueden destacar las siguientes conclusiones y recomendaciones, útiles tanto para futuras investigaciones como para aplicaciones en estudios de ciencia abierta y evaluación de la producción científica.

El repositorio de OpenAlex brinda más resultados (campos, etiquetas, y/o columnas), que los expuestos en Crossref. Sin embargo, Crossref también es un repositorio grande, por lo que este hecho puede estar relacionado con el funcionamiento de la API de Crossref o, también, con la consulta realizada a partir del paquete `rcrossref`. Además, el formato de datos (matriz) que devuelve, es menos manejable en comparación con el formato entregado por Openalex (`openalexR`). La principal ventaja de `openalexR` en relación con `rcrossref` es que permite buscar por un campo de identificación como el `Id_openalex` o `Ror_Id`. Haciendo que la localización de datos sea menos complicada. A diferencia de `rcrossref` que se tuvo que incluir las variaciones de nombres de las universidades para obtener los datos. Cuando se intentó buscar por un campo de identificación `cr_works(filter=c(funder="funder_Id")` en `rcrossref`, este devolvía pocos resultados.

La comparación con WoS, la cual es una base con criterios de indexación más estrictos, permitió validar y/o contrastar los datos obtenidos de repositorios abiertos. Esto aporta solidez a los resultados y evidencia la importancia de combinar fuentes de distinta naturaleza para obtener un panorama más completo y confiable de la producción científica. Una extensión podría ser el de considerar todas las colecciones que WoS alberga, y compararlas con las de OpenAlex. Las colecciones que se consideraron aquí fueron SCI, SSCI y A&HCI. Por otro lado, también se podría ampliar el estudio con otras bases de datos como Scopus y/o Dimensions. Igualmente, se podría desarrollar páginas interactivas (dashboards o informes) con `RShiny`(libre) o `PowerBi`(comercial), que expongan los resultados de manera más amigable con los usuarios, similar a `biblioshiny()`, solo que esta debería tener más opciones o menús desplegables.

Se realizó un análisis general de las áreas de estudio que se han desarrollado dentro del SUG, respecto a un periodo de tiempo único (rango de años), una posible ampliación podría darse al estudiar el cambio dinámico por año que tienen las áreas de estudio, y así ver una evolución en las áreas de estudio para el SUG.

Por otro lado, se podría tratar de crear una función similar a `oa2df()` de `openalexR`, para el caso de `rcrossref`, de modo que los datos extraídos puedan estructurarse de forma uniforme y sean fácilmente procesables por herramientas como `bibliometrix` o `scimetr`. Esta estandarización permitiría aplicar métodos bibliométricos más avanzados, comparativos y/o reproducibles.

Al parecer la librería `rcrossref`, no permite acceder completamente a los registros de la API de Crossref, tanto en cantidad, como en algunas variables de los registros, por ejemplo no se hallan: CR, JI, J9, DI (variables no tan relevantes). Entre las variables relevantes que no devuelve están `Concepts` o `Topics` las cuales permitirían realizar un análisis de co-ocurrecia de palabras, similar a como se haría en OpenaAlex o Wos. Por lo que

se podrían explorar otras librerías con otros lenguajes, para ver si estas tienen las mismas limitaciones que `rcrossref` u `openalexR`, por ejemplo las librerías: `habanero` y `crossrefapi` de `Phyton`.

Se podrían plantear estudios de simulación para probar aspectos técnicos de las librerías como: velocidad, conexión, descarga, capacidad, etc. Así como también se podrían considerar otros análisis cualitativos de estadística multivariante, por ejemplo, un análisis de correspondencias o técnicas de agrupación (clústeres), que identificarían posibles grupos o patrones presentes en variables como: `tittle`, `abstract`, etc.

Bibliografía

Aguillo-Caño, I. F. (2012). *Is Google Scholar useful for bibliometrics? A webometric analysis*. Scientometrics, 91(2):343-351.

Aksnes, D. W., Langfeldt, L. y Wouters, P. (2019). *Citations, Citation Indicators, and Research Quality: An Overview of Basic Concepts and Theories*. SAGE Open, 9(1):1-17.

Ardanuy, J. (2012). *Breve introducción a la bibliometría*. Universitat de Barcelona, Departament de Biblioteconomia i Documentació. URL: https://diposit.ub.edu/dspace/handle/2445/30262. (Visitado 20 de mayo de 2025).

Aria, M. y Cuccurullo, C. (2017). *bibliometrix: An R-tool for comprehensive science mapping analysis*. Journal of Informetrics, 11(4):959-975.

Aria, M., Le, T., Cuccurullo, C., Belfiore, A. y Choe, J. (2024). *OpenAlexR: An R-Tool for Collecting Bibliometric Data from OpenAlex*. The R Journal, 15(4):167-180.

Arruda, H., Silva, E. R., Lessa, M., Proença Jr, D. y Bartholo, R. (2022). *VOSviewer and bibliometrix*. Journal of the Medical Library Association: JMLA, 110(3):392-392.

Baas, J., Schotten, M., Plume, A., Côté, G. y Karimi, R. (2020). *Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies*. Quantitative Science Studies, 1(1):377-386.

Borrego-Huerta, Á. (2024). *OpenAlex: breve guía de consulta a través de R*. Infonomy, 2(1):1-8.

Chamberlain, S., Zhu, H., Jahn, N., Boettiger, C. y Ram, K. (2014). *Rcrossref: Client for Various "CrossRef" "APIs"*. The R Foundation. URL: http://dx.doi.org/10.32614/cran.package.rcrossref. (Visitado 25 de junio de 2025).

Cooper, D. (2015). *Bibliometrics basics*. Journal of the Medical Library Association: JMLA, 103(4):217.

Cuadras-Avellana, C. M. (2007). *Análisis de correspondencias*. En: Cuadras-Avellana, C. M. (ed.), Nuevos métodos de análisis multivariante, CMC Editions, Barcelona, pp. 161-182.

Drongstrup, D. (2021). *Bibliometrics and Co-Authorship*. En: Strotmann, A. y Horstmann, W. (eds.), Handbook Bibliometrics, De Gruyter Saur, Berlin, Boston, pp. 397-406.

Eck-Nees, J. v. y Waltman, L. (2022). *Crossref as a source of open bibliographic metadata*. MetaArXiv. URL: https://osf.io/preprints/metaarxiv/smxe5/. (Visitado 3 de febrero de 2025).

Fernandez-Casal, R., Lafuente-Rego, B., Lombardia, M. J., Tarrio-Saavedra, J. y Costa-Bouzas, J. (2023). *Scimetr: Analysis of Scientific Publication Data with R*. The R Foundation. URL: https://github.com/rubenfcasal/scimetr. (Visitado 25 de junio de 2025).

Fernández-Ramos, Y. (2021). *Análisis y Diseño de un Sistema de Información Bibliométrico de Publicaciones Científicas*. bachelor's thesis, Universidade da Coruña.

Gingras, Y. (2016). *Bibliometrics and Research Evaluation: Uses and Abuses*. MIT Press, Cambridge, MA.

Glänzel, W. (2003). *Historical Remarks*. En: KU Leuven (ed.), Bibliometrics as a Research Field: A Course on Theory and Application of Bibliometric Indicators, Course Handouts, Leuven, pp. 6-9.

Hirsch, J. E. (2005). *An Index to Quantify an Individual's Scientific Research Output*. Proceedings of the National Academy of Sciences, 102(46):16569-16572.

Kleminski, R., Kazienko, P. y Kajdanowicz, T. (2022). *Analysis of Direct Citation, Co-Citation and Bibliographic Coupling in Scientific Topic Identification*. Journal of Information Science, 48(3):349-373.

Krampl, A. (2019). *Journal Citation Reports*. Journal of the Medical Library Association, 107(2):280-283.

Leung, X. Y., Sun, J. y Bai, B. (2017). *Bibliometrics of Social Media Research: A Co-Citation and Co-Word Analysis*. International Journal of Hospitality Management, 66:35-45.

Malvić, T., Andreić, Ž., Barudžija, U., Bedeković, G., Hrnčević, L., Ivšinović, J., Korman, T., Kovač, Z., Pavlić, K. y Pašić, B. (2022). *Citation Rate Challenges for a Small Journal Indexed in Scopus and WoS—Case Study from Central Europe (Croatia), Editorial View*. Publications, 10(3):32.

Martínez-Reglero, C. (2018). *Análisis de datos espaciales de devoluciones*. Master's thesis, Universidad de Santiago de Compostela.

Mongeon, P. y Paul-Hus, A. (2016). *The Journal Coverage of Web of Science and Scopus: A Comparative Analysis*. Scientometrics, 106(1):213-228.

Pranckutė, R. (2021). *Web of Science (WoS) and Scopus: The titans of bibliographic information in today's academic world*. Publications, 9(1):12.

Rubio-Liniers, M. C. (1999). *Bibliometría y ciencias sociales*. AIBDA, 20(7):138-155.

Tomás-Górriz, V. y Tomás-Casterá, V. (2018). *La Bibliometría en la evaluación de la actividad científica*. Hospital a domicilio, 2(4):145-163.

Touriñan-López, J. M. (2025). *El sistema universitario de Galicia. Consolidación, desarrollo y retos (una aproximación 35 años después)*. Revista Redipe, 14:41-72.

Vence-Deza, X. (2010). *Sistema gallego de innovación y políticas de innovación: Balance de una década*. Revista Galega de Economía, 19(1):1-30.

Waltman, L. (2016). *A review of the literature on citation impact indicators*. Journal of Informetrics, 10(2):365-391.

Siglas y Glosario

API

Application Programming Interfaces. 3, 6, 8, 17–19, 21, 24, 25, 27, 33, 62

BBDD

Bases de Datos. 2, 3, 6, 8, 17, 19, 21, 24, 28, 33, 36, 40, 41, 43, 48, 62

Bibliometrix

Paquete o librería de ámbito estadístico dentro del repositorio de R, el cual se orienta a la investigación cuantitativa en cienciometría y bibliometría, para mayor información véase la documentación oficial de esta librería. . 17, 23, 28, 62

Crossref

Asociación abierta sin fines lucro, orientada a buscar una eficiencia en términos de registros para las publicaciones académicas, gestionando y proveyendo de la infraestructura digital necesaria para la comunidad académica en todo el mundo. 2, 3, 6, 8, 18, 19, 23, 25, 28, 33, 36, 39, 40, 62

DOI

Digital Object Identifier, Referente a una clave única o identificador, que tiene como objetivo, identificar los documentos digitales (libros, tesis, artículos, etc), de manera única . 3, 6, 7, 18, 62

ISSN-L

International Standard Serial Number Linking, Identificador o clave única, con el cual se puede identificar a un conjunto o familia de publicaciones seriadas, que están en diferentes formatos, tiene como objetivo vincular los formatos existentes de una misma publicación (seriada), bajo un identificador único . 3, 18, 62

OAID

Open Archives Initiative Identifier, Similar a DOI, al tambien ser único, permite la navegación en varios repositorios académicos, permitiendo que sea distinguible en los repositorios . 3, 62

OpenAlex

Conjunto de datos académicos de código abierto en linea, que cuenta con una estructura o interfaz de búsqueda, para poder acceder a datos de los entes académicos tales como: autores, conceptos, instituciones, entre otros. 2, 3, 6, 8, 17, 19, 23, 24, 28, 33, 36, 38, 40, 62

ORCID

Open Researcher and Contributor ID, Identificador o clave única (16 dígitos en formato estándar de URI), con el cual se puede identificar a un autor (evitando así, problemas de traducción de nombres, nombres dobles, entre otros), independiente de su lugar de trabajo . 3, 6, 7, 62

ROR ID

Research Organization Registry Identifier, Identificador o clave única (formato de URL), con el cual se puede identificar a instituciones, universidades u organizaciones de investigación en todo el mundo . 3, 24, 25, 62

Scimetr

Paquete o librería de ámbito estadístico dentro del repositorio de R, similar a bibliometrix, aunque en desarrollo, permite obtener índices bibliométricos y representaciones generales dentro de la bibliometría, para mayor información véase la documentación oficial de esta librería.. 17, 28, 62

SUG

Sistema Universitario de Galicia. 23–25, 33, 36, 46, 48, 50, 53, 62

UDC

Universidad de A Coruña. 23, 24, 33, 36, 40, 41, 43, 44, 46, 48, 50, 62

USC

Universidad de Santiago de Compostela. 23–26, 33, 36, 40, 41, 43, 46, 48, 51, 62

UVigo

Universidad de Vigo. 23, 24, 33, 36, 40, 41, 43, 46, 48, 51, 62

Wikidata ID

Identificador o clave única, para identificar elementos que se encuentran en el repositorio de Wikidata, organiza la información de manera estructurada. Es importante tener en cuenta que en este ámbito, los conceptos siguen una jerarquía y permiten clasificar de mejor manera una palabra, por ejemplo:

- Nivel 0: Ciencias de la Computación Nivel 1: Inteligencia Artificial.
- Nivel 2: Aprendizaje Automático Nivel 3: Redes Neuronales Artificiales.
- Sucesivamente hasta llegar al concepto o tema necesario.

. 3, 7, 62

WoS

Web of Science. 2, 3, 6, 18–21, 40, 62

Apéndice

Script para OpenAlex

```
# Ventana/Script Openalex
# Librerías.....................................................#
library(openalexR); library(tidyverse); library(bibliometrix)
library(ggraph); library(igraph); library(tidygraph)
library(FactoMineR); library(factoextra)

# Función para consultar y obtener datos.........................#
Getoplx <- function(Id_openalex,Ror_Id){

  query <- oa_query(entity = "works",from_publication_date = "2014-01-01",
             to_publication_date = "2024-12-31", verbose = TRUE,
  authorships.institutions.id=Id_openalex, authorships.institutions.ror=Ror_Id)
  res <- oa_request(query_url = query,count_only = FALSE,verbose = FALSE)
  df <- oa2df(res, entity = "works"); M <- oa2bibliometrix(df); return(M)}
# ..............................#
# Consutas para cada universidad del SUG
# Nota: El tiempo de consulta es algo lento, debido a la cantidad y dependencia
#       de la API, por ello, se recuerre a guardar las bases en memoria local,
#       para luego cargarlas en memoria con load().

Udc_M_Opnlx <- Getoplx(Id_openalex="I11019714",  Ror_Id ="01qckj285")
Usc_M_Opnlx <- Getoplx(Id_openalex="I200284239", Ror_Id ="030eybx10")
Udv_M_Opnlx <- Getoplx(Id_openalex="I6289922",   Ror_Id ="05rdf8595")

# save(Udc_M_Opnlx,file="Udc_M_Opnlx.Rda"); save(Usc_M_Opnlx,file="Usc_M_Opnlx.Rda")
# save(Udv_M_Opnlx,file="Udv_M_Opnlx.Rda")
#.................................................................#
# load("Udc_M_Opnlx.Rda"); load("Usc_M_Opnlx.Rda"); load("Udv_M_Opnlx.Rda")
#.................................................................#
# Función para Crear matriz similar formato "WoS" (reetiquetado en Openalex)
# Ya no sería necesaria esta función debido a que no sabía de la existencia
# de oa2df(), el cual adapta el formato similar a WoS, aunque  la dejo para
# probarla con Scimetr, aun así no se adaptaba la librería a los datos, en fin
# la dejé por si adelante servía.

CreatedDB.opnlx <- function(matrix){
  matrix %>% rowwise() %>% mutate(
    AU = paste(author$au_display_name, collapse = "; "), # Autores
      TI = title,  # Título
      SO = so,  # Fuente
      PY = publication_year,  # Año de publicación
```

```
      DI = doi,  # Doi
      VL = volume,  # Volumen
      IS = issue,  # Número de la revista
      PP = ifelse(first_page == last_page, first_page, paste0(first_page, "-",
                  last_page)),  # páginas
      AB = ab,  # Resumen
      TC = cited_by_count,  # Número de citas
      CR = paste(referenced_works, collapse = "; "), # Referencias citadas
      UT = id,  # Identificador (id de OpenAlex)
      DB = "OpenAlex",  # Base de datos
      DB_ID = id  # ID en OpenAlex
    ) %>% ungroup() %>%
    select(TI, AU, SO, PY, DI, VL, DB, IS, PP, AB, TC, CR, UT, DB_ID) }

# Matriz OpenAlex para el SUG
data_sug_opnlx <- bind_rows(Udc_M_Opnlx,Usc_M_Opnlx,Udv_M_Opnlx)
data_sug_opnlx$SUG <- factor(rep(c(1,2,3), times = c(nrow(Udc_M_Opnlx),
                                nrow(Usc_M_Opnlx), nrow(Udv_M_Opnlx))))
levels(data_sug_opnlx$SUG)=c("1"="UDC","2"="USC","3"="UDV")
data_sug_opnlx  <- data_sug_opnlx %>% filter(PY %in% 2014:2024)
```

Script para CrossRef

```
library(rcrossref); library(tidyverse); library(lubridate)
library(openxlsx); library(bibliometrix); library(ggplot2)
#.....................................................................#
# Función general para importar trabajos desde Crossref
Getcrf <- function(affiliation) {
  all_results <- list(); limit <- 950 ; cursor <- "*" ;
  while (!is.null(cursor)) {
    tryCatch({
      result <- cr_works( filter = c( from_pub_date = "2014-01-01",
          until_pub_date = "2024-12-31"),
        query = paste0(affiliation, collapse = " OR "),
        cursor = cursor, limit = limit)
      if (!is.null(result$data)) {
        all_results <- append(all_results, list(result$data))}
      cursor <- result$meta$`next-cursor`
      if (is.null(cursor)) {break}
      Sys.sleep(1)}, error = function(e) {
      cat("Se produjo un error: ", e$message, "\n")
      cursor <- NULL }) }
  results <- do.call(rbind, all_results)
  return(results) }

# Afiliaciónes, nombres
names_udc <- c(
"Universidade da Coruña", "University of A Coruña","University of A Coruna",
"Universidad de La Coruña", "Universidad de A Coruña", "Univ. Coruña",
"Univ Coruña", "UDC", "Coruña University", "University of Corunna",
"University of Coruña", "Universidade da Corunha")
works_udc_crf <- Getcrf(names_udc)%>%
          filter(purrr::map_lgl(author, ~ is.data.frame(.x) && nrow(.x)>0))
```

```r
names_usc <- c(
"Universidade de Santiago de Compostela",
"University of Santiago de Compostela", "USC", "Univ. Santiago de Compostela",
"Univ Santiago de Compostela", "Santiago de Compostela University")
works_usc_crf <- Getcrf(names_usc)%>%
          filter(purrr::map_lgl(author, ~ is.data.frame(.x) && nrow(.x)>0))
#.....................................................................#
names_udv <- c("Universidade de Vigo", "University of Vigo", "UVigo",
                "Univ. Vigo", "Vigo University")
works_udv_crf <- Getcrf(names_udv)%>%
          filter(purrr::map_lgl(author, ~ is.data.frame(.x) && nrow(.x)>0))
#.....................................................................#
# Nota: La estructura y variables de las matrices anteriores imposibilitan utilizar
# directamente funciones como Bibliometrix o Scimetr, por ello se necesitará:
# Transformación de datos a formato WoS
#.....................................................................#

# Función para referencias anidadas (tipo .Json)

format_ref <- function(ref) {
       if (is.null(ref) || nrow(ref) == 0) { return(NA_character_)}
  ref_strings <- ref %>% mutate(
      author = ifelse("author" %in% names(.), author, NA),
      year = ifelse("year" %in% names(.), year, NA),
    article.title = ifelse("article.title" %in% names(.), article.title, NA),
    journal.title = ifelse("journal.title" %in% names(.), journal.title, NA),
      volume = ifelse("volume" %in% names(.), volume, NA),
      first.page = ifelse("first.page" %in% names(.), first.page, NA),
      DOI = ifelse("DOI" %in% names(.), DOI, NA)) %>%
    pmap_chr(~ paste(na.omit(c(...)), collapse = ", "))
  paste(ref_strings, collapse = "; ") }

# Función extraer/formatear "SR" desde reference(matriz de works_udc_crf)

extract_SR <- function(ref_list) {
  if (is.null(ref_list) || !is.data.frame(ref_list) || nrow(ref_list) == 0) {
    return(NA_character_)  # Si es NULL o no tiene filas, devuelve NA
    }
  # Extrae y une las referencias si la columna 'unstructured' existe
  if ("unstructured" %in% names(ref_list)) {
    return(paste(na.omit(ref_list$unstructured), collapse = "; ")) }

  return(NA_character_)  # Si no hay 'unstructured', devuelve NA
}

# Funcion Principal

CreatedDB.crf <- function(matrix){
  matrix%>%rowwise()%>% mutate(
    AU = paste(author$given, author$family, sep = " ", collapse = ";"),
           TI = title,
           SO = container.title,
           PY = year(as.Date(created)),
           DI = doi,
           VL = volume,
           IS = issue ,
```

```
          PP = ifelse(!is.na(page), page, NA) ,
          AB = abstract ,
          TC = as.integer(is.referenced.by.count) ,

    CR = if (!is.null(reference) &&
                  is.data.frame(reference)) format_ref(reference) else NA,

          UT = doi,
          DB = "CrossRef",
          SR=extract_SR(reference),
          SN=issn,
          EI=issued, # Opcionales para Scimetr
          NR=as.numeric(reference.count),
          SN=issn,
          LA=language,

          PG= ifelse(grepl("-", page),as.numeric(sub("-.*","",page)
                    )-as.numeric(sub(".*-", "", page)), NA),

          DT=type,
          PU=publisher,
          PD=toupper(format(as.Date(created), "%b%d"))
      ) %>%
    ungroup()%>% select(AU,TI,SO,PY,DI,VL,IS,PP,AB,TC,CR,UT,DB,funder,SN,EI,
                        NR,SN,LA,PU,PD,DT,PG) }

#.....................................................................#

# Matriz Crossref para el SUG
data_sug_crf <- bind_rows(CreatedDB.crf(matrix = works_udc_crf),
                          CreatedDB.crf(matrix = works_usc_crf),
                          CreatedDB.crf(matrix = works_udv_crf) )
data_sug_crf$SUG <- factor(rep(c(1,2,3), times = c(nrow(works_udc_crf),
                        nrow(works_usc_crf),nrow(works_udv_crf))))
levels(data_sug_crf$SUG)=c("1"="UDC","2"="USC","3"="UDV")
data_sug_crf <- data_sug_crf %>% filter(PY %in% 2014:2024)
```

Script para WoS

```
# Decargados los arcchivos WoS (Web), importese a R mediante:
#.....................................................................#
import <- function(directorio) {
archivos <- list.files(directorio,pattern="\\.txt$",full.names=TRUE)
# Convertir cada archivo a un dataframe
lista_df <- lapply(archivos, function(directorioe) {
convert2df(directorioe, dbsource = "wos", format = "plaintext") })
# Campos/columnas en común
columnas_comunes <- Reduce(intersect, lapply(lista_df, names))
lista_directoriotrada <- lapply(lista_df,function(df)df[,columnas_comunes])
return(do.call(rbind, lista_directoriotrada)) }
#.....................................................................#
udc_M_wos <- import(directorio = "C:/Users/...")
usc_M_wos <- import(directorio = "C:/Users/...")
udv_M_wos <- import(directorio = "C:/Users/...")
```

```r
#.....................................................................#
# Matriz wos para el SUG
data_sug_wos <- bind_rows(udc_M_wos,usc_M_wos,udv_M_wos)
data_sug_wos$SUG <- factor(rep(c(1,2,3), times = c(nrow(udc_M_wos),
                         nrow(usc_M_wos),nrow(udv_M_wos))))
levels(data_sug_wos$SUG)=c("1"="UDC","2"="USC","3"="UDV")

data_sug_wos <- data_sug_wos %>% filter(PY %in% 2014:2024) %>% mutate(
      DT_aux = case_when(
# Articulos-Review(y variantes)
DT %in% c("ARTICLE", "ARTICLE; BOOK CHAPTER",
"ARTICLE; DATA PAPER","ARTICLE; EARLY ACCESS", "ARTICLE; PROCEEDINGS PAPER",
"ARTICLE; PUBLICATION WITH EXPRESSION OF CONCERN",
"ARTICLE; RETRACTED PUBLICATION",
"ARTICLE; WITHDRAWN PUBLICATION","REVIEW", "REVIEW; BOOK CHAPTER",
"REVIEW; EARLY ACCESS","REVIEW; RETRACTED PUBLICATION")~"Articulo o revision",

# Libros y capítulos (variantes de BOOK)
DT %in% c("BOOK","BOOK CHAPTER","BOOK REVIEW",
          "BOOK REVIEW; EARLY ACCESS")~"Libro o Capítulo",

# Material editorial y cartas
DT %in% c("EDITORIAL MATERIAL", "EDITORIAL MATERIAL; BOOK CHAPTER",
"EDITORIAL MATERIAL; EARLY ACCESS","LETTER",
      "LETTER; EARLY ACCESS")~ "No evaluable",

# Correcciones y retracciones
DT %in% c("CORRECTION","CORRECTION; EARLY ACCESS",
          "RETRACTION")~"No evaluable",

# Difusión sin revisión por pares
DT %in% c("MEETING ABSTRACT","PREPRINT","DATASET")~"Difusión sin revisión",

# Otros tipos diversos
DT %in% c("NEWS ITEM","BIBLIOGRAPHY","BIOGRAPHICAL-ITEM",
          "SOFTWARE REVIEW")~"Otro",
# Casos especiales
DT %in% c("EXHIBIT REVIEW","POETRY","REPRINT") ~ "Otro",
# Por defecto
  TRUE ~ "Otro"))
```