Universidad de Salamanca

Departamento de Informática y Automática

Evaluación de la producción científica mediante motores de búsqueda académicos y de acceso libre.

Tesis doctoral presentada por:

D. José Federico Medrano

Directores:

Dr. D. José Luis Alonso Berrocal

Dr. D. Carlos G. Figuerola

Junio 2017

UNIVERSIDAD DE SALAMANCA
Departamento de Informática y Automática

Evaluación de la producción científica mediante motores de búsqueda académicos y de acceso libre.

Tesis doctoral presentada por:
D. José Federico Medrano

Dirigida por:
Dr. D. José Luis Alonso Berrocal
Dr. D. Carlos G. Figuerola

El doctorando

Salamanca, Junio de 2017

Dr. D. José Luis Alonso Berrocal y Dr. D. Carlos G. Figuerola , *Profesores Titulares del Departamento de Informática y Automática de la Universidad de Salamanca*

HACEN CONSTAR: *Que* D. José Federico Medrano, *Máster en Sistemas Inteligentes por la Universidad de Salamanca ha realizado bajo nuestra dirección la Memoria que lleva por título* "Evaluación de la producción científica mediante motores de búsqueda académicos y de acceso libre", *con el fin de obtener el grado de Doctor por la Universidad de Salamanca.*

Y para que surta los efectos oportunos firmamos en Salamanca, a nueve de Junio de dos mil diecisiete.

A mi hija ***Victoria****,*

con todo mi amor y con cada

átomo de mi cuerpo,

eres mi luz, mi sentido y

mi más grande motivación.

Agradecimientos

En primer lugar quiero agradecer a mis tutores José Luis y Carlos. Primero por haberme brindado su confianza y por haber apostado en mí. Segundo, por haber sembrado la idea desde el inicio y haber cuidado el proceso completo con una guía y dedicación que pocos la tienen. Tercero por haber sido mi motivación desde lejos, sobre todo cuando aún faltaba muchísimo por hacer no dejaron de creer. Y por último por haber despertado en mí la admiración primero hacia sus personas y segundo hacia la pasión que le ponen cada día a su trabajo, si estas personas no se habrían cruzado en mi camino no habría descubierto mi pasión por la recuperación y visualización de información. Sinceramente y sencillamente gracias.

En segundo lugar quiero agradecer a la Fundación Carolina por haberme becado para iniciar mis estudios de máster y por haber confiado en mí la gran responsabilidad de aprender. El haberme permitido cruzar el charco, no sólo expandió mis conocimientos y mi sed de superación, sino que me hizo descubrir un mundo lleno de posibilidades. A veces, este tipo de instituciones no mide el impacto de sus pequeñas acciones ni cómo eso puede cambiar la vida de una persona. Hoy puedo decir que la inversión que hicieron en mi dio sus frutos.

Un agradecimiento enorme a la Universidad de Salamanca, en primer lugar por abrirme sus puertas y presentarme desafíos que creía imposibles de superar, segundo por haber confiado en mis posibilidades y otorgarme una beca para iniciar mis estudios de doctorado, en tercer lugar por tener a personas tan grandes como docentes, capaces no sólo de enseñar e impartir conocimientos, sino también capaces de dejar sus huellas y marcar el rumbo a sus estudiantes.

También quiero agradecer a mi familia, a mis padres sobre todo, son

dos personas que con estudios básicos fueron capaces de insistir en mi formación permanente, que siempre estuvieron presentes, apoyando y alentando mis locuras y mis sueños. Cuando el camino se vuelve cuesta arriba, es imposible subir sin que alguien empuje desde atrás, gracias mamá. Y en especial quiero agradecer a mi mujer y mis hijos, por el apoyo, la paciencia y por aguantar mis ausencias y mi mal humor, sobre todo en la etapa final.

Quiero agradecer a todas las personas que tuve el placer de conocer y que de una forma u otra ayudaron, colaboraron y motivaron mis ganas de ser hoy un poco mejor que ayer. No solo se aprende de las cosas buenas de la vida, de las malas y de los errores, también.

Este trabajo sin confianza y sin apoyo no sería una realidad. Este proyecto se realizó haciéndome cargo de mi familia, teniendo dos trabajos a tiempo completo y con muy pocas horas de sueño. Tuve que renunciar a mucho para lograr este objetivo, y no ha sido un camino de rosas, pues entiendo que "*quien no ha afrontado la adversidad no conoce su propia fuerza (Benjamin Jonson)*". No conozco otra forma de hacer las cosas que no sea dando el cien por ciento.

Y un último agradecimiento a mi hija Victoria por ser mi más bella inspiración, se que no podrá leer estas líneas pues solo tiene un año y medio, pero esto es por ti y esto es para ti mi niña. Ojalá, de alguna forma, algún día, tu padre pueda ser un ejemplo para ti, al menos en el corazón y alma que dejé para hacer realidad mis sueños... *Vicky*, si se puede.

To give anything

less than your best

is to sacrifice the gift

Steve Prefontaine (1951-1975)

Resumen

La evaluación de la producción científica o específicamente la evaluación de la productividad de un científico, ha sido desde que se iniciaran las publicaciones de los resultados de la investigación, una tarea nada sencilla. El ser humano en su naturaleza de cuantificarlo todo, ha desarrollado a lo largo del tiempo, variadas métricas y metodologías para contabilizar los frutos del trabajo de la investigación y comunicarlo a la comunidad científica. Cualquier tarea que implique una medida lleva consigo el empleo de un conjunto de técnicas, herramientas, fórmulas y reglas para asegurar la objetividad de los resultados arrojados.

Para obtener indicadores de productividad es necesario tener acceso a los datos que servirán para el análisis, en este caso, acceso a los resultados de la investigación en forma de publicaciones científico-académicas. Las bases de datos tradicionales como lo son *Scopus* y *Web of Science* han sido los referentes en este ámbito por muchísimo tiempo, pero en la última década y media se ha comenzado a gestar, y en los últimos años con más fuerza, la apertura de esta información, tal es el caso de la existencia de los motores de libre acceso, el ejemplo favorito de este tipo de motores es sin lugar a dudas *Google Scholar*, aunque *Microsoft* ha realizado una apuesta firme con la nueva versión de su motor académico *Microsoft Academic*.

La existencia de motores académicos de libre acceso ha abierto la puerta a un sin número de estudios encargados de evaluar dichas herramientas desde distintos puntos de vista: la cobertura y la autenticidad de la información son dos de los más populares. Estos motores han evolucionado al punto que pasaron de ser simples interfaces de búsqueda de material académico a ser consideradas herramientas de evaluación. A raíz de esto se generó un debate interesante entre los encargados de realizar análisis bibliométricos, pues dichos motores al ser de libre acceso

presentan ciertos problemas como la falta de normalización de los datos, problema del que no escapan incluso las bases de datos tradicionales.

La falta de normalización en cualquier tipo de bases de datos no es un problema trivial y si sumado a ello, estas bases de datos son utilizadas para evaluar la productividad de un académico, para otorgar subvenciones, becas o promover un ascenso, el problema se torna un tanto complejo. No se debe perder de vista que dicho inconveniente desencadena una serie de problemas no menores como la existencia de duplicados, la imposibilidad de identificar claramente homónimos, la existencia de material no científico-académico, entre muchos otros.

Hoy por hoy *Google Scholar* presenta algunos indicadores bibliométricos en su motor académico con lo cual la discusión se amplía más aun, pues dichos indicadores se basan en la información que recolecta este motor, es más, empresas encargadas de realizar análisis bibliométricos toman como fuente de datos los provistos por este buscador.

Habiendo dado un pequeño panorama de la situación actual, esta tesis pretende ofrecer una alternativa a los resultados que ofrecen este tipo de motores académicos de libre acceso, pues se entiende que el principal problema es la falta de normalización. Es así, que en un intento de mejorar la calidad de la información brindada, en este trabajo se desarrolló una herramienta que toma como origen de datos las publicaciones de un científico que se pueden recoger de *Google Scholar*, *Microsoft Academic* o cualquier otro motor académico. La idea es procesar estos datos poco normalizados y obtener indicadores más cercanos a la realidad, o al menos libres de los inconvenientes mencionados.

Para ello, en primer lugar se ofrece un estado de la cuestión enfocado en los problemas que se han recogido de la bibliografía existente, también se realiza un estudio de los indicadores y de las formas más comúnmente utilizadas a la hora de evaluar la producción científica, se presenta un pantallazo de las herramientas que actualmente evalúan o recogen de otras herramientas información para evaluar el estado actual de la productividad de un académico, analizando el conjunto de publicaciones como es el esquema habitual y analizando también el impacto de estas en las redes sociales tanto científico-académicas como redes sociales comunes (alternativas *altmetrics*), por último en esta primera parte se revisan los esquemas, técnicas y herramientas más importantes utilizados en la actualidad para resolver estos problemas y las posibilidades en cuanto a

visualización de información que ofrecen las bases de datos actuales.

Luego de la parte introductoria, se describen todos los procesos y mecanismos generados primero para recuperar la información de estos motores y luego para procesar dicha información. El procesamiento que se dio a los datos se divide en tres partes, la primera de ellas permite desambiguar los autores de un conjunto de publicaciones, este primer proceso evalúa los patrones de colaboración que se establecen entre los coautores de una publicación, esto permite agrupar conjuntos de autores que al parecer se conocen o colaboran entre sí, solo se cuenta con los datos ofrecidos por una publicación de un motor poco normalizado, no se cuenta con datos adicionales para resolver este problema, es la incertidumbre quien guía el proceso de análisis y es en base a esta información que se realizan las inferencias para obtener un resultado aproximado. La segunda parte del proceso se encarga de desambiguar los registros de las publicaciones de estos grupos formados, para ello detecta los posibles duplicados y realiza una fusión de las citas recibidas por estos registros. La tercera parte del proceso ofrece los resultados obtenidos primero como un conjunto de indicadores y luego como una visualización interactiva desarrollada totalmente para este trabajo, que logra exhibir algunas de las dimensiones y variables involucradas al momento de evaluar un conjunto de publicaciones.

Una vez detallado el modelo que se diseñó para procesar la información, se realiza un experimento real para comprender no sólo como funciona la herramienta sino también para entender el procesamiento completo llevado a cabo y tener así una imagen completa de la complejidad de esta tarea.

Por último se resumen los resultados obtenidos, resaltando las mejoras y ventajas que aporta la solución planteada, como el hecho de intentar resolver la ambigüedad en nombres de autores identificando todas las posibles variantes y excepciones en la forma y modo de firmar de un autor, resuelve también la ambigüedad de los títulos de los registros de publicaciones, parte de un conjunto de datos poco normalizado y realiza un ajuste y limpieza adecuado entregando indicadores más reales, se vale del uso de herramientas diseñadas para propósitos académicos como la *Academic Knowledge API* o de búsquedas de propósito general como la *Bing Web Search API* de *Microsoft*, además propone una visualización interactiva que mejora en gran medida las visualizaciones básicas y estáticas ofrecidas por las bases de datos actuales.

Pese a las ventajas del modelo expuesto, este presenta ciertas limitaciones como la utilización de una única fuente de datos al mismo tiempo, la demora en el procesamiento para algunos casos, pues ante la falta de datos y de información para el proceso de desambiguación, es necesario consultar información de la web y obtener datos que den algún indicio de las relaciones entre dos grupos de autores al parecer no relacionados, además algunos de los procesos corren en línea, con lo cual la demora es real pues no se realiza un procesamiento inicial al almacenar los datos, como lo realizan otras bases de datos. De un modo u otro, estos problemas se podrán resolver en líneas futuras de investigación ampliando los límites de los orígenes de datos, empleando otras técnicas y mecanismos que aumenten el conocimiento inicial, por ejemplo incluyendo datos adicionales como datos filiatorios, datos geográficos, información del campo de estudio, técnicas de PLN entre otros.

Abstract

The evaluation of scientific production or, more specifically the evaluation of the productivity of a scientist, ever since the start of the publishing of research results, has not been a simple task. The human being, in his nature of quantifying everything, has developed through time varied metrics and methodologies to count the fruits of the research work and communicating it to the scientific community. Any task that carries a measure brings a set of techniques, tools, formulas and rules to assure the objectivity of the given results.

To obtain productivity indicators we need to access the data that will be used in the analysis, which in this case is access to the results of the research in the form of a scientific/academic publication. The traditional databases, such as Scopus and Web of Science, have been the leaders for a long time, but in the last decade and a half, and especially in the last years, this information is becoming more open, as is the case with free access search engines. The favorite example for these search engines is Google Scholar without any doubt, but Microsoft has done a firm bet with the new version of their academic engine, Microsoft Academic.

The existence of free access academic engines has open the door to countless studies in charge of evaluating those tools from different points of view: coverage and authenticity of the information are two of the more popular. These engines have evolved to the point in which they have gone from simple search interfaces for academic material, to be considered evaluation tools. This has generated an interesting debate between the people in charge of doing bibliometric analysis, because those engines, being free to access, present certain problems, as the lack of normalization of the data, a problem that not even traditional databases can escape from.

The lack of normalization in any kind of database is not a trivial problem. And if adding to this, these databases are used to evaluate the productivity of an academic, to give grants, scholarships or a promotion, the problem becomes a bit complex. We should not forget that these inconveniences trigger a series of non-minor problems, such as the existence of duplicates, the impossibility to clearly identify homonymous, the existence of non-scientific/academic material, among many others.

Nowadays Google Scholar presents some bibliometric indicators in their academic engine, making the discussion even broader, because those indicators are based in the information collected by the engine. Even more, companies in charge of doing bibliometric analysis take the data given by this search engine as a source.

Having given a short status of the current situation, this thesis wants to offer an alternative to the results these kind of free access academic engines offer, because we understand that the main issue is the lack of normalization. Thus, in an attempt to improve the quality of the provided information, this work has developed a tool that uses the publications of a scientist that can be collected from Google Scholar, Microsoft Academic or any other academic engine as a source. The idea is to process these hardly normalized data and to obtain indicators that are closer to reality, or at least free from the aforementioned inconveniences.

To do this, firstly we offer a state of the issue focused on the problems that have been gathered from existing bibliography. We also make a study of the indicators and the most commonly used ways of evaluating scientific production. We present a snapshot of the tools that currently evaluate or collect information from other tools to evaluate the current state of the productivity of an academic, analyzing the set of publications with the usual schema, and also analyzing their impact in social media, both scientific/academic as common social media networks (altmetrics alternatives). Lastly, this first part revises the most important schemas, techniques, and tools currently used to solve these problems and the possibilities of information visualization current databases offer.

After the introduction part, we describe all the processes and mechanisms generated firstly to gather the information from these engines and then to process this information. The processing used with the data is divided in three parts. The first one allows us to disambiguate the authors from a set of publications. This first process evaluates the collaboration

patterns established by coauthors of a publication. This allows us to group sets of authors that seem to know or collaborate with each other. We only count on the data offered by a publication a poorly normalized engine. We do not have any additional data to solve this problem, it is uncertainty that guides the analysis process, and the inferences to obtain an approximate result are based on this information. The second part of the process deals with the disambiguation of the records of the publications from these formed groups. To do this it detects possible duplicates and does a fusion of the citations gotten by these registries. The third part of the process offers the obtained results, first as a set of indicators, and then as an interactive visualization, developed entirely for this work, that gets to show some of the dimensions and variables involved at the moment of evaluating a set of publications.

Once the model designed to process the information is detailed, a real experiment is conducted to comprehend not only how the tool works, but also to understand the complete processing that was executed and thus have a complete image of the complexity of this task.

Lastly, the obtained results are summarized, highlighting the improvements and advantages the given solution presents, such as the fact of trying to solve the ambiguity in author's names, identifying all the possible variations and exceptions in the signature and the way of signing of the author. It also solves the ambiguity in the titles of the publication records. It starts from a poorly normalized set of data, and it makes the adjustment and cleaning, delivering more real indicators. It makes use of tools designed for academic purposes such as the Academic Knowledge API, or general purpose searches such as Microsoft's Bing Web Search API. It also proposes an interactive visualization that improves in great measure the basic and static visualizations offered by current databases.

Despite the advantages of the exposed model, it presents certain limitations as the use of a single source of data at the same time, the delay in the processing for certain cases, because of the lack of data and information for the disambiguation process, it is necessary to query information from the web and to get data that give a hint of the relationships between two groups of authors that appear not to be related. Besides, some of the process run online, so the delay is real because there is no initial processing to store the data, as other databases do. One way or the other, these problems can be solved in future lines of research broadening the limits of the data sources, using other techniques and

mechanisms that increase the initial knowledge, for example, including additional data such as filial data, geographical data, information from the field of study, and PNL techniques, among others.

Índice general

Índice de tablas

Índice de figuras

Índice de Algoritmos

Capítulo 1

Introducción y objetivos

1.1. Introducción

En estos últimos años evaluar y cuantificar la productividad científica se ha convertido en tema de un profundo análisis ya que las cantidades de información aumentan constantemente y se vuelven cada vez más dinámicas. En este momento existen variados métodos cuantitativos para medir esta producción, entre los más conocidos y usados destacan: el factor de impacto y h-*index*. Si bien estos números reflejan la relación entre artículos publicados y la popularidad de los mismos, entiéndase esto en términos de la cantidad de veces que un artículo es citado por otro o la importancia asignada a la revista en donde se publique, muy a menudo no resultan suficientes o precisos. Es importante notar que no es suficiente medir cuán a menudo un recurso es utilizado, sino más importante es considerar cómo es utilizado. Por ello el análisis basado en las citas bibliográficas es una forma muy común de evaluar estos factores y es quizás el modo más adecuado en este ámbito.

Dos de las herramientas existentes en el mercado, entre las más importantes y reconocidas, que se encargan de medir la producción científica son: *Scopus (Elsevier)* y *Web of Science (ISI/Thomson)*. Si bien estas herramientas son fiables ya que los resultados entregados provienen de fuentes confiables y comprobadas, poseen ciertas limitaciones a la hora de presentar los resultados de forma global. Por ejemplo no tienen en cuenta los libros, ni otros materiales publicados en revistas de menor impacto o de otra índole como tesis, informes u otros documentos.

En este sentido, motores de búsqueda académicos como *Google Scholar (GS)* o *Microsoft Academic Search (MAS)* se presentan como una buena alternativa por varias razones (otros motores de búsqueda académicos de libre acceso son *CiteSeerX* y *DBLP*):

- El rango de cobertura es amplio, no solo artículos de revistas y congresos sino libros, tesis, informes y muchos otros tipos de documentos.
- Son herramientas de libre acceso.
- No están circunscritos a determinadas áreas del conocimiento, y cada vez el alcance es mayor hacia áreas de menor relevancia.
- Son muy populares y de fácil uso.
- Son útiles para cualquier trabajo académico sin importar la envergadura de este.

Sin embargo no todo es bueno para este tipo de herramientas de libre acceso, muchas veces los resultados entregados no son muy confiables o a veces no son los esperados, ya que en el caso de GS incluye un bajo control de la calidad. Una parte muy importante de este problema responde a la falta de normalización (ambigüedad) en los nombres de autores y en los títulos de los documentos. El problema de la ambigüedad se presenta no solo en este tipo de bases de datos, sino en la mayoría de las fuentes de datos, ya que muchas veces provienen de errores humanos (error de escritura, traducción, al no firmar un documento de la forma habitual, etc.).

Una cuestión que ha sido un poco descuidada en este ámbito y sobre todo en estas herramientas (sin importar si son de libre acceso o no), es la visualización de la información. Estas herramientas se limitan a presentar simples estadísticas o datos tabulados, o los conocidos grafos de co-autor o co-citas, dejando de lado algún otro tipo de visualización más compleja como la evolución temporal involucrando varias dimensiones más que la cantidad de citas por año.

Por todos estos motivos, la idea central de este trabajo es la creación de una herramienta que mida la productividad científica utilizando como fuente de datos motores de libre acceso como *Google Scholar* y *Microsoft*

Academic, y resolviendo los problemas de ambigüedad tanto en los nombres de los autores referenciados como en los títulos de las publicaciones, eliminando las citas duplicadas y descartando los resultados incorrectos. Así mismo se pretende encontrar la mejor forma, la más adecuada o proponer una nueva visualización para representar las relaciones existentes entre autores, publicaciones, citas, y cualquier otro tipo de información o relación que se considere relevante durante el estudio.

1.2. Objetivos

El objetivo principal de ésta tesis es intentar dar solución a un gran problema: lograr medir cuán productivo es un científico tomando conjuntos de publicaciones de fuentes poco normalizadas, obteniendo indicadores de productividad, y representaciones de los datos y variables involucradas.

De este problema se derivan varias cuestiones a resolver, por un lado la existencia de publicaciones duplicadas (derivadas de diferentes fuentes de datos), registros mal indizados por el motor (porque entiéndase que tanto *Google Scholar* como *Microsoft Academic*, no son generadores de contenidos sino indizadores de contenidos publicados en sitios web), la existencia de autores homónimos (más de un autor que posee exactamente el mismo nombre o parte del nombre del autor buscado), las variantes de nombres de un mismo autor (a veces un autor no firma de la misma forma o el repositorio lo almacena incorrectamente o simplemente el motor lo indiza incorrectamente).

El método desarrollado debe contemplar los siguientes objetivos globales:

1. El primero de ellos es crear una herramienta que permita reunir los resultados de dos motores académicos de libre acceso: *Google Scholar* y/o *Microsoft Academic*. Para ello se creará un *crawler* que sea capaz de consultar y procesar en tiempo real las consultas realizadas por el usuario contra estos motores desde una única interfaz.

2. El segundo objetivo es resolver el problema de la ambigüedad de los nombres de autores y de los títulos de las publicaciones. Como se dijo anteriormente, este problema se presenta en la mayoría de

las fuentes de datos y es una cuestión muy importante a resolver a la hora de determinar la productividad de un determinado autor, grupo o institución científica. En la bibliografía se pueden encontrar diversos trabajos abordando el tema desde distintos puntos de vistas y alternativas para dar solución a los problemas particulares que se estudian, es decir, hoy en día no existe una solución absoluta o genérica para las diferentes variantes de este problema, por ello en este trabajo se estudiarán las distintas variantes y se implementará aquella que mejor se adapte y que entregue un buen balance entre tiempo de ejecución y efectividad. Es necesario destacar que en este trabajo el tiempo de proceso/respuesta es muy importante ya que los resultados y el procesamiento se obtienen en tiempo real.

3. Por último, el objetivo final es diseñar una visualización novedosa, ágil e interactiva para el usuario, que involucre la mayor cantidad de dimensiones posibles de los datos extraídos y procesados (autores, número de publicaciones, número de citas). En este sentido se dejarán de lado los simples gráficos estadísticos presentados hasta el momento.

1.3. Motivación

La motivación para llevar a cabo este trabajo resulta clara pues a día de hoy no existe una herramienta que trabaje sobre estos motores académicos y resuelva de forma integral las variedad de problemas presentados: extracción de información, desambiguación de nombres, detección de duplicados y visualización de información. Si bien existen numerosos trabajos en la bibliografía que han realizado alguna aproximación a estos problemas, siempre resulta limitada y acotada a un conjunto definido de datos.

Algunas de las soluciones estudiadas y analizadas, se restringen a una base de datos de alguna área temática, o tratando solo un problema al mismo tiempo y no todo el espectro, o trabajando sobre un conjunto de datos controlados (entiéndase por controlado, datos que no cambian con el tiempo o provenientes de fuentes de datos normalizadas).

Google Scholar ha sido el objeto de estudio de un gran número de trabajos, es muy criticado por algunos autores pero también es muy reconocido por la mayoría, además de ser no solo mundialmente conocido

y de libre acceso, sino también capaz de cubrir una amplia gama de contenido (por no decir todo). Por su parte *Microsoft Academic*, luego de haber resurgido hace un año, ha logrado una cobertura más que notable, superando con creces a *Scopus* y *Web of Science*, y estando solo unos pasos por detrás de GS, por estos motivos, estos dos motores (GS y MA) han sido seleccionados como fuentes de datos para los experimentos y para amoldar y ajustar el modelo de tratamiento de la información que se ha diseñado en este proyecto.

Si bien con esta tesis no se intenta dar una solución exacta o definitiva a estos problemas, sino una aproximación razonable que contemple todas las variantes e implicaciones de dichos problemas, de manera de poder contar con una herramienta útil y sencilla a la hora de cuantificar o medir la productividad científico/académica de un investigador.

Por otro lado, intentar plasmar las distintas variables de este proceso en una visualización que vaya más allá de un gráfico estadístico o datos tabulados, para poder interpretar la mayor cantidad de dimensiones involucradas en este tema (basada principalmente en la visión del autor), resulta más que atractivo, ya que hasta el momento lo existente se limita a la utilización de gráficos de barras, tablas o grafos (en el mejor de los casos).

1.4. Alcance y limitaciones

El alcance del presente plan de trabajo contempla lo que se fijó en la sección de objetivos. Se debe aclarar que para éste trabajo se tendrá en cuenta la productividad de un autor objeto de la búsqueda, y no de un grupo de investigación o institución, ya que los algoritmos se adecuarán para tratar autores individuales.

Se tomará como fuente y origen de datos la información proporcionada por cualquier motor académico de libre acceso, ejemplo de ellos son *Google Scholar* o *Microsoft Academic*, o bases de datos tradicionales como *Scopus* o *Web of Science*, o cualquier otro motor ya que la información a procesar será importada en un formato específico que será detallado en el **Capítulo 3**. Sólo para *Microsoft Academic* se podrá realizar la recolección en linea, ya que dicho motor posee una API para tal fin.

Si bien el alcance de esta tesis está demarcado en los objetivos planteados, hay que remarcar que el ámbito es interdisciplinario, puesto que se nutre de un conjunto de disciplinas para lograr estos objetivos de los que se habla. Haciendo una analogía al modo de definir el alcance del trabajo desarrollado por (Pascual Cid, 2010), en la Figura 1.1 se puede ver el alcance de esta tesis.

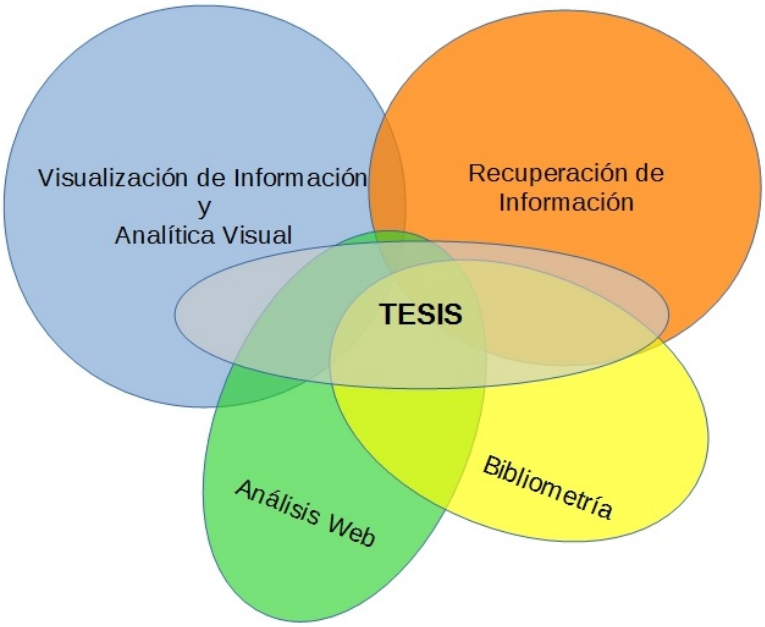


Figura 1.1: Alcance de la tesis

Por lo tanto en esta tesis se estudiará, desarrollará y evaluará un modelo de procesamiento que permitirá tener una opción al momento de evaluar fuentes de datos poco controladas o faltas de normalización para medir la productividad de un científico, para ello se involucrará tareas de las siguientes disciplinas que a lo largo de la redacción se verá la implicación de cada una de ellas: Visualización de Información y Analítica Visual, Recuperación de Información, Análisis Web y Bibliometría.

1.5. Organización de la Tesis

Esta tesis inicia con una breve introducción del problema a tratar, indica cual es la motivación real y los objetivos para el desarrollo de este trabajo de investigación. Luego la tesis se estructura de la siguiente manera:

- En el capítulo 2 se presenta una descripción de los principales indicadores utilizados para medir la productividad científica, aquellos propios de bases de datos bibliográficas tradicionales, como los de uso universal. También se describe brevemente algunas de las herramientas de libre acceso para obtener información académica e indicadores de productividad provenientes de motores de búsqueda académicos. Luego se presenta un estudio y clasificación tanto de los métodos utilizados como de los trabajos realizados para resolver los problemas de desambiguación de autores y detección de duplicados. Por último, en este capítulo se realiza una introducción a las técnicas de visualización de información, y a las posibilidades de visualización que ofrecen algunas de las herramientas y bases de datos bibliográficas mencionadas anteriormente.
- En el capítulo 3 se exponen los detalles de las técnicas, métodos y herramientas utilizados en este trabajo para resolver los problemas mencionados. Se hace una descripción de los algoritmos empleados y diseñados para este trabajo, así como la justificación de los indicadores utilizados para cuantificar la productividad científica. Por último, en este capítulo se describe la técnica de visualización empleada para plasmar los resultados del proceso.
- El capítulo 4 detalla la arquitectura y las opciones de la aplicación desarrollada para dar soporte a todos los procesos diseñados.
- En el capítulo 5 se presentan los resultados de los experimentos realizados con la herramienta construida, la comparación con algunas herramientas disponibles, y las visualizaciones obtenidas de estas pruebas.
- Por último, el capítulo 6 muestra las conclusiones derivadas de esta tesis, así como las líneas de trabajos futuros.

Capítulo 2

Estado del arte

2.1. Introducción

En este capítulo se discute el estado actual de las áreas de estudio relacionadas con esta investigación: Productividad Científica, Recuperación de Información, Desambiguación y Visualización de Información relacionados con la evaluación y análisis bibliométricos.

2.2. La productividad científica

La investigación científica se inicia con la búsqueda de información, su consumo racional y la elaboración de un proyecto que finalizará con la difusión de sus resultados y hallazgos a través de la publicación de un documento. Este documento se diseminará a través de artículos publicados en revistas, capítulos de libros, ponencias presentadas en congresos u otros tipos de documentos similares (Restrepo Arango and Urbizagástegui Alvarado, 2010). De esta forma es cómo la ciencia y los resultados de las investigaciones realizados por los científicos se dan a conocer y no existen solo como una práctica individual o limitada a pequeños grupos.

Según (Sancho, 1990) "... ha surgido la necesidad de evaluar el rendimiento de la actividad científica y su impacto en la sociedad con el fin primordial de adecuar convenientemente la asignación de los recursos destinados a investigación y desarrollo, punto indispensable en la gestión

y planificación científica de cualquier institución o país para conseguir una rentabilidad máxima en las inversiones en este campo". Partiendo de la necesidad de contabilizar este esfuerzo, se empezó a poner énfasis en controlar o medir en cierta forma la productividad científico-académica. Es decir, teniendo en cuenta la cantidad de elementos que son publicados.

Al hablar de contabilizar se tendría que tener en cuenta dos aspectos muy importantes: la cantidad y la calidad. Que un investigador o grupo de investigadores produzcan gran cantidad de material no quiere decir que necesariamente sea relevante o importante, lo que se llama visibilidad de un trabajo. Además, existen campos del conocimiento donde se publica muy poco o donde es muy difícil tener un volumen importante de publicaciones.

Como indica (Urbizagástegui Alvarado, 2005), esta tendencia ha dado lugar a reflexiones intelectuales sobre lo que ahora es considerado como la obligación de publicar y la existencia de un grupo de significantes contribuyentes en cualquier campo del conocimiento. Por lo tanto, se podría preguntar si la contribución de los grandes productores es de menor, igual, o mayor calidad que la contribución de los menores productores.

Los estudios de la productividad científica se iniciaron en la década de los 20's, con los estudios de (Dresden, 1922) sobre la producción de artículos de autores ligados a la Sociedad Americana de Matemáticas, cuatro años más tarde, (Lotka, 1926; Gorbea Portal, 2005) propuso la ley del cuadrado inverso o también conocida como ley de Lotka, para medir la productividad de los autores en un campo científico.

La Ley de Lotka expresa que el número de autores que hacen $\boldsymbol{n}$ contribuciones es aproximadamente $\frac{1}{n^2}$ de aquellos que solo hacen una contribución, es decir, independientemente de la disciplina el número de autores que publican $\boldsymbol{n}$ trabajos es inversamente proporcional a n^2.

La ley del cuadrado inverso de la productividad científica tiene la siguiente fórmula matemática:

$$A_n = \frac{A_1}{n^2}$$

Donde:

- A_n es el número de autores con $\boldsymbol{n}$ trabajos
- A_1 es el número de autores con 1 trabajo
- n^2 es el número de trabajos al cuadrado.

Sus resultados constataron que la producción científica es variable de acuerdo al campo de investigación que, por lo que solo unas pocas personas contribuyen al progreso de la ciencia en gran medida, mientras que la mayoría contribuye muy poco. La aplicación de su ley puede usarse para saber con qué frecuencia publica un autor y cuál es la relevancia de sus trabajos, aunque debe advertirse que la productividad de los científicos no tiene que coincidir necesariamente con la calidad de sus trabajos.

Estos estudios no solo se centraron en medir la cantidad de publicaciones, sino en los factores o variables que influyen en dicha producción, tal es el caso de (Fox, 1983) que indica entre las variables que influyen en este número a los rasgos psicológicos, hábitos de trabajo, características demográficas como la edad, contexto social y cultural y prestigio de la institución. Por su parte, (Simonton, 1999) analiza la productividad desde la perspectiva psicológica a través del estudio de la creatividad de un investigador. Los sociólogos, (Merton, 1977; Allison and Steward, 1974), buscan una explicación de la productividad científica a través de la teoría de la ventaja acumulativa, quien más produce posee mayor reconocimiento y prestigio que quien produce menos.

2.2.1. Bases de datos bibliográficas

Las grandes bases de datos bibliográficas como lo son *Web of Science (WoS)* (ISI/Thomson) [1] y *Scopus* (Elsevier)[2], han sido durante mucho tiempo el elemento para medir la producción científica de los investigadores y la visibilidad de sus trabajos. *WoS* creada por Thomson Reuters (que forma parte del *Institute for Scientific Information ISI*) es una base de datos bibliográfica comercial online (contiene el texto completo y resúmenes de artículos) accesible a través del portal de *Web of Knowledge (WoK)*[3] y ofrece acceso a cinco bases de datos de citas completas:*Science*

[1] http://thomsonreuters.com/en/products-services/scholarly-scientific-research/scholarly-search-and-discovery/web-of-science.html

[2] http://www.scopus.com

[3] http://apps.webofknowledge.com/

Citation Index Expanded, Social Science Citation Index, Arts & Humanities Citation Index, Book Citation Index, Conference Proceedings Citation Index, Current Chemical Reactions, Index Chemicus y Emerging Sources Citation Index [4].

En el año 2012, Thomson Reuters lanzó el *Data Citation Index (DCI)* como "Un único punto de acceso a datos de investigación de calidad, provenientes de múltiples repositorios de tres de las principales areas del conocimiento (Ciencia y Tecnología, Ciencias Sociales, y Artes y Humanidades) y del mundo entero" (Pavlech, 2016; Torres-Salinas et al., 2014b). El DCI incluye tres tipos diferentes de documentos: *datasets* (conjunto de datos), *data studies* (estudios de datos) y *repositories* (repositorios)(Torres-Salinas et al., 2014a).

Los "repositorios" de datos se definen como bases de datos que almacenan y proporcionan acceso a los datos brutos contenidos en los conjuntos de datos y los estudios de datos. Los "conjuntos de datos" son un conjunto único y coherente de datos proporcionados como parte de una colección, de estudios de datos o experimentos en uno o más archivos. Finalmente, los "estudios de datos" se definen como una descripción de los experimentos con los datos asociados que se han utilizado en estos experimentos (Robinson-Garcia et al., 2015). Esta herramienta solo se encuentra disponible para uso exclusivo con licencias institucionales a través de la plataforma de *Web of Science*. En (Torres-Salinas et al., 2014b) se ofrece un estudio cuantitativo de la cobertura del DCI mientras que en (Torres-Salinas et al., 2014a) queda evidenciado la reducida cantidad de citas que existen en la herramienta y por consiguiente la alta tasa de documentos no citados (un importante 88 %).

La *WoS* es la opción de búsqueda y exploración de más de 7.000 instituciones académicas y de investigación, gobiernos nacionales, organizaciones de financiación y organizaciones de editoriales en más de 100 países al rededor de todo el mundo. Abarca revistas científicas, libros, actas, conjuntos de datos publicados, y patentes. Este contenido es verdaderamente global y multidisciplinario (más de 250 disciplinas), procedente de 80 países diferentes, y en 32 idiomas. La *WoS Core Collection* posee mas de 12.500 revistas de alto impacto, más de 170.000 actas de congresos, más de 70.000 libros (Reuters, 2016c).

Scopus se lanzó en Noviembre de 2004, una iniciativa creada por

[4] http://wokinfo.com/products_tools/multidisciplinary/webofscience/

Elsevier, es una lista bibliográfica comercial, que a diferencia de *WoS*, es más amplia en lo que a lengua y países de publicación se refiere y es hoy por hoy el principal competidor de *WoS*. *Scopus* es la mayor base de datos de resúmenes y citas de material de investigación revisado por pares (peer-reviewed) con más de 60 millones de registros. Incluye al rededor de 21.500 revistas *peer-reviewed* (con 4.200 revistas de Acceso Abierto), más de 5.000 editores internacionales, también incluye libros: mas de 130.000 y 10.000 agregados cada año. Las áreas de mayor cobertura, como lo indica la Figura 2.1 son Medicina, Física, Ciencias Sociales y Ciencias de la Vida(Elsevier, 2016b).

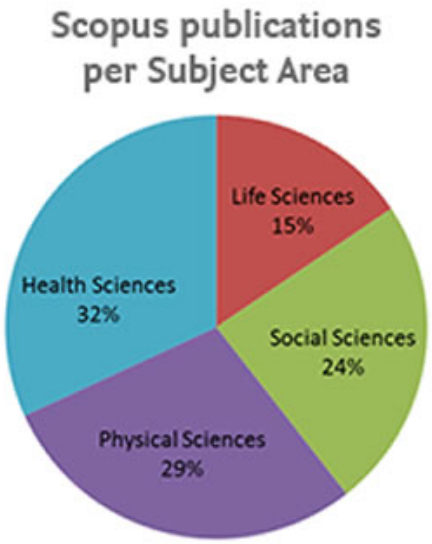


Figura 2.1: Porcentaje de publicaciones por área temática en Scopus

Estas bases de datos poseen mecanismos de control que aseguran que los datos y los índices ofrecidos son correctos, aun así la cobertura no es la deseada ya que solamente se encuentran trabajos de revistas con un factor de alto impacto, es decir, revistas de bajo impacto u otros materiales académicos, como libros, no son tenidos en cuenta. Además, estas bases de datos no son públicas, para tener acceso a ellas es necesario abonar una suscripción (a veces solo alcanzable por grandes corporaciones e instituciones). En estos últimos años han cobrado relevancia las

bibliotecas digitales de libre acceso tales como *Google Scholar* (GS) [5], Microsoft Academic Search (MAS)[6], DBLP [7] y citepSeerX [8] que contienen grandes cantidades de registros de citas bibliográficas de artículos académicos, libros, tesis, documentos no oficiales, entre otros, y en muchos casos permiten el acceso al texto completo del documento en cuestión. Estas librerías se han convertido en un recurso importante para la comunidad académica ya que permiten búsquedas de referencias y materiales relevantes a la hora de encarar un trabajo a cualquier escala.

Comparar las bases de datos bibliográficas tradicionales (*WoS* y *Scopus*) frente a las bases de datos de libre acceso (*Google Scholar* y/o *Microsoft Academic*) resulta una tarea no del todo sencilla, particularmente por la cobertura de éstas. Como se mencionó, las bases de datos tradicionales indexan material puramente académico y de un grupo determinado de revistas y congresos, es un ambiente controlado y el margen de error (si lo hubiera) es mínimo, además no todas las áreas del conocimiento son indexadas por este tipo de base de datos.

La *WoS* indexa el contenido de las siguientes áreas temáticas: Ciencias de la Vida y Biomedicina, Física, Tecnología, Artes y Humanidades, y Ciencias Sociales(Reuters, 2016b); por su parte Scopus ofrece a los investigadores un recurso rápido, fácil y completo para apoyar sus necesidades de investigación en todos los campos de investigación de la Ciencia, las Matemáticas, la Ingeniería, la Tecnología, la Salud y la Medicina, las Ciencias Sociales, y las Artes y Humanidades(Elsevier, 2016b).

Google Scholar en cambio indexa todo lo que puede, no solo lo relativo al tipo de publicación (artículos académicos, libros, tesis, tesinas, memorias de grado, actas de congresos, resúmenes, pre-impresos, y todo material que aparente tener contenido académico que cuelgue de un dominio académico) sino cualquier área temática mientras la información del recurso esté disponible, esto no quiere decir que el acceso al recurso completo este permitido, porque puede que el acceso esté restringido pero la información de dicho recurso sí está disponible y accesible a este tipo de rastreadores.

En lo que a cobertura se refiere también hay que tener en cuenta el idioma de la publicación. Las publicaciones en idiomas distintos al

[5] http://scholar.google.com
[6] http://academic.research.microsoft.com/default.aspx
[7] http://www.dblp.org/search/index.php
[8] http://citepseerx.ist.psu.edu/index

Inglés sufren un sesgo importante ya que las grandes bases de datos bibliográficas indexan solo una parte de estas obligando en cierta forma a utilizar el Inglés como lengua de difusión de sus trabajos.

En un análisis de (Santa and Herrero-Solana, 2010) sobre la cobertura de la ciencia de América Latina y el Caribe en *Scopus* vs *Web of Science*, sobre datos recogidos en 2008, indica la mayor cobertura de *Scopus* (utilizando el *Scimago Journal and Country Rank - SJR*) frente a *Web of Science* (utilizando la fuente de *Journal Citation Report - JCR* disponible en WoK), SJR indexa 444 revistas producidas en América Latina y el Caribe frente a las 79 indexadas por JCR. En cuanto al lenguaje de publicación se ve cómo en SJR tienen mayor predominio los títulos en español que suponen el 35 % del total, frente al 20 % que representan en JCR, al igual que las revistas en portugués, que alcanzan el 27 %, frente al 10 % de JCR, con las publicaciones en inglés sucede lo contrario, ya que es en JCR donde éstos tienen mayor peso 30 % frente a 12 % en SJR.

Un estudio de (Leydesdorff et al., 2010) sobre la cobertura del idioma de las revistas indexadas por el ISI y Scopus desde 1996 hasta 2007, refleja la amplia cobertura de Scopus de revistas publicadas en idiomas distintos del Inglés y menos populares como Turco, Húngaro, Eslovaco y Polaco entre otros.

En un trabajo de (Harzing and van der Wal, 2009), se evidencia la falta de inclusión de revistas en lenguajes distintos del Inglés en las bases de *WoS*. Por ello las revistas que no publican artículos en Inglés probablemente reciban menor cantidad de citas debido a que una gran parte de la comunidad científica no puede o no lee otros idiomas (Adler et al., 2009). Sin embargo, comparando la cobertura de *Google Scholar* frente a las otras bases de datos, se puede ver que éste ofrece una cobertura significativamente mayor de materiales en idiomas distintos al inglés.

(Meho and Yang, 2007) en un estudio sobre el impacto de las citas almacenadas en las bases de datos de *WoS*, *Scopus* y GS sobre los trabajos de 25 miembros de la facultad de Biblioteconomía y Documentación entre 1996-2005, encontró, en cuanto a la distribución de la cantidad de citas por idioma del material publicado en idiomas distintos del Inglés que GS cubre el 6,94 % del total de citas, frente al 1,14 % de *WoS* y 0,70 % Scopus. Esto sugiere que *Google Scholar* es indispensable para mostrar el impacto de revistas y artículos en idiomas distintos al Inglés.

Por otro lado, en (Harzing and Alakangas, 2016) basados en una

muestra de 146 académicos de alto nivel en cinco disciplinas principales, se presenta una comparación longitudinal y transversal a través de las 3 bases de datos más importantes: Google Scholar, Scopus y Web of Science. El análisis longitudinal mostró un crecimiento trimestral consistente y razonablemente estable tanto para las publicaciones como para las citas recibidas, en las tres bases de datos. Esto sugiere que las tres bases de datos proporcionan suficiente estabilidad de la cobertura que se utilizará para las comparaciones transversales. Luego, la comparación transversal de las métricas utilizadas indicó que tanto la fuente de datos como los indicadores específicos utilizados cambian las conclusiones que pueden extraerse de las comparaciones transversales.

Una investigación comparativa reciente de (Mongeon and Paul-Hus, 2016), describe la cobertura de las revistas tanto de *Wos* y *Scopus*, y evalúa si algún campo, país e idioma de publicación son sobre o subrepresentados. Para ello evaluaron la cobertura de las revistas científicas vigentes (13.605 revistas) y *Scopus* (20.346 revistas), utilizando el directorio periódico de *Ulrich* (63.013 revistas) [9].

Cada una de estas revistas cubre de forma diferente cada campo del conocimiento, por consiguiente, el uso de cualquiera de ellas (*WoS* o *Scopus*) para la evaluación de la investigación puede introducir sesgos que favorecen a las Ciencias Naturales e Ingeniería, así como a la investigación Biomédica en detrimento de las Ciencias Sociales, y Artes y Humanidades. Del mismo modo, las revistas en idioma Inglés están sobrerrepresentadas en detrimento de otros idiomas. Si bien ambas bases de datos comparten estos sesgos, su cobertura difiere sustancialmente. Como consecuencia, los resultados de los análisis bibliométricos pueden variar en función de la base de datos utilizada.

Estos resultados implican que en el contexto comparativo de evaluación de la investigación, *WoS* y *Scopus* se deben utilizar con precaución, especialmente cuando se comparan diferentes campos, instituciones, países o idiomas. También hay discusiones sobre la exactitud, precisión y cobertura de *Wos* y *Scopus*, como lo indican los trabajos de (Wang and Waltman, 2016; Waltman, 2016; Aghaei Chadegani et al., 2013) las comparaciones entre estas bases de datos se basan normalmente en la evaluación de los indicadores bibliométricos calculados por cada una de ellas, la visibilidad, la cantidad y calidad de las revistas que indexan, los áreas

[9] http://ulrichsweb.serialssolutions.com/

y categorías que resultan mejor cubiertas por una u otra, entre otras cuestiones.

Existen numerosos trabajos encargados de analizar la cobertura de *Google Scholar* frente a otras bases de datos bibliográficas (Moed et al., 2016; Groote and Raszewski, 2012; Clermont and Dyckhoff, 2012; Walters, 2011; Miguel and Solana, 2010; Torres-Salinas et al., 2009; AV et al., 2009; Bar-Ilan, 2008; Yang and Meho, 2006), la mayoría coincide que si bien *Google Scholar* no indexa todo el universo posible (mucho del contenido oficial no es de libre acceso con lo cual no está disponible para ser rastreado por los robots de *Google*), sí posee en relación a las áreas de interés y al tipo de publicaciones una mayor cobertura que *WoS* y *Scopus*, en la Figura 2.2 se puede ver el solapamiento de las citas proporcionadas por *Google Scholar*, *Scopus* y *WoS* en el campo de la documentación, esta figura fue tomada del trabajo de (Torres-Salinas et al., 2009).

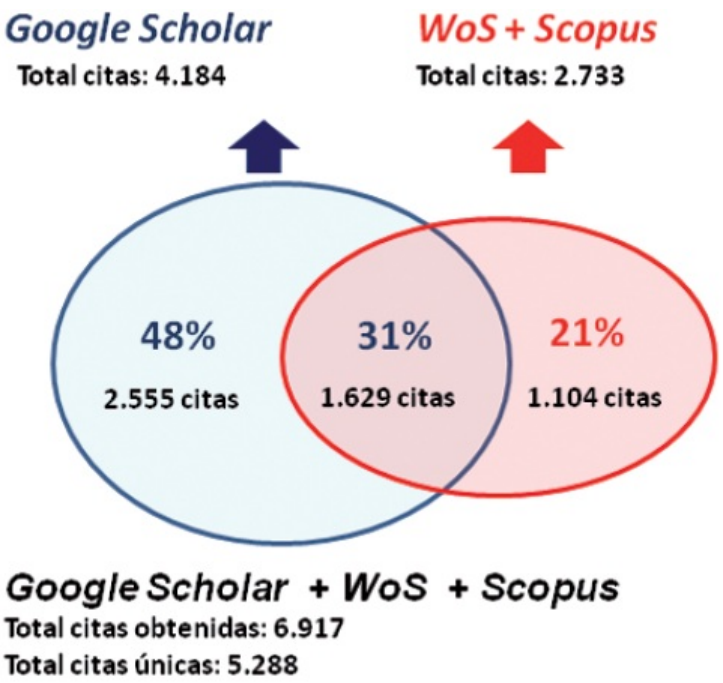


Figura 2.2: Ejemplo del solapamiento de las citas proporcionadas por GS, *Scopus* y *WoS* en el campo de la documentación

2.3. Análisis de citas, indicadores

El crecimiento de la producción científica en las últimas décadas así como su recopilación en bases de datos bibliográficas automatizadas ha potenciado el uso de la "Bibliometria" (Bibliometrics) y la generación de indicadores para medir los resultados de la actividad científica y tecnológica. Los indicadores bibliométricos son datos estadísticos deducidos de las distintas características de las publicaciones científicas, en base al importante papel que desempeñan estas en la difusión y transmisión del conocimiento generado en la investigación.

Son válidos cuando los resultados de la investigación se transmiten a través de publicaciones científicas y técnicas. Proporcionan información cuantitativa y objetiva sobre los resultados del proceso investigador, su volumen, evolución, visibilidad y estructura, pero no informan sobre los progresos del conocimiento(Mainardi and de Morán-Suárez, 2011).

Una investigación profunda y exhaustiva es llevada a cabo por (Waltman, 2016) sobre los indicadores de impacto de citas, tomando como fuente de datos *Wos*, *Scopus* y GS. Este estudio analiza los indicadores más importantes relacionados con el recuento de citas de publicaciones, teniendo en cuenta las diferencias que existen al comparar estos indicadores entre campos distintos, indicando que para ello es necesario un proceso de normalización de datos, ya que reconoce que existen áreas donde la publicación de material es mucho más frecuente que en otras, por ejemplo, un bioquímico con 25 citas no puede considerarse que posee un impacto de citas mayor que un matemático con 10 citas. Lo mismo sucede con el año de publicación, existen casos en que hay años más prolíficos que otros, y en el mismo sentido el tipo de publicación, sea que se trata de un artículo, una revisión o libro.

Hoy en día el análisis o recuento de citas bibliográficas es el mecanismo más efectivo o al menos el más utilizado para medir la productividad científico-académica, no solo por su sencillez de aplicación sino por su transparencia. Constituye un pilar a la hora de conceder subvenciones y financiamientos de proyectos y en la toma de decisiones, ya que son cada vez más utilizados como una medida de desempeño por el cual los científicos y profesores se clasifican, son promovidos y financiados, además el análisis de citas permite a los investigadores dar seguimiento al desarrollo e impacto de un artículo a través del tiempo.

Según (Capaccioni and Spina, 2012), "La cuestión más debatida es, por supuesto, cuáles son los criterios de evaluación. ¿Criterios únicos o criterios diferentes para las diferentes áreas? ¿Criterios cualitativos o criterios cuantitativos?". La importancia sobre la evaluación de la investigación científica se ha incrementado consistentemente en la primera década del nuevo siglo, sumado a la aplicación de algoritmos sofisticados que actualmente se encuentran disponibles para el análisis de citas permitiendo evaluar la "calidad" de estas, permitió la proliferación de un gran número de indicadores, entre ellos dos grupos bien diferenciados, los que se encargan de evaluar una revista científica y los que evalúan la productividad de un científico.

Existen otros factores que deben tenerse en cuenta en la medición e interpretación de las estadísticas, para asegurarse de que las comparaciones entre los recursos son relevantes, por ejemplo, la "edad" de la publicación puede afectar notablemente la tasa de uso, y existen grandes diferencias en cómo las publicaciones son utilizadas y accedidas a través del tiempo.

2.3.1. Impact Factor

El *Impact Factor - IF*(Garfield, 1972, 2006) o también conocido como *Journal Impact Factor - JIF*, propuesto por *Eugene Garfield* fundador del *Institute for Scientific Information (ISI)*, es un instrumento para comparar revistas científicas y para evaluar la importancia relativa de una revista dentro de un mismo campo científico, es probablemente el indicador bibliométrico más utilizado en la comunidad científica, académica y editorial. Representa un promedio derivado de la distribución de citas para una colección de artículos publicados en la revista.

Si bien útil, este indicador debe ser utilizado con mucha precaución como lo indican (Stonebraker et al., 2012; Adler et al., 2009), no es una herramienta infalible ya que para obtener una correcta evaluación es necesario tener muy en cuenta el contexto, el área de estudio, la ventana de aplicación (2 años es lo común, pero también se pueden aplicar 5 años), y muy importante qué es lo que se está evaluando si una revista o un artículo.

Como lo indica (Falagas et al., 2008), aunque el IF ha sido ampliamente considerado como el mejor instrumento para la evaluación de la

calidad de las revistas científicas, no se ha librado de las críticas. Los principales puntos de consideración son sobre aspectos metodológicos en el cálculo de este índice, como la falta de evaluación de la calidad de las citas, la inclusión de auto-citas, la escasa comparabilidad entre los distintos campos científicos, y sobre todo el análisis de publicaciones principalmente en idioma Inglés.

2.3.2. Journal Citation Reports

Las bases bibliográficas tradicionales ofrecen indicadores para evaluar las revistas incluidas en sus bases de datos, al tener el material disponible en estas y al poseer las bases de datos de citas, resulta una tarea bastante sencilla. En lo relativo a *WoS* existe el *Journal Citation Reports (JCR)*, que recopila las referencias de los artículos citados, permitiendo medir la influencia de la investigación y el impacto a nivel de revistas y categorías, y muestra la relación entre la cita y las revistas citadas, permite además comparar revistas entre sí.

JCR reúne a más de 10.100 revistas de más de 238 disciplinas, publicados en más de 2.600 editores de 84 países. JCR se encuentra disponible en dos ediciones: *JCR Science Edition* que cubre alrededor de 8.000 revistas, y *JCR Social Science Edition* con 2650 revistas, citando estadísticas de 1997 [10]. El *JCR* proporciona herramientas cuantitativas para la clasificación, evaluación, categorización y comparación de revistas. El *IF* es uno de éstos, el cual se calcula como el cociente entre el número de citas y los últimos artículos citables que han sido publicados. Por lo tanto, el *IF* de una revista se calcula dividiendo el número de citas del año en curso entre los elementos publicados en esa revista durante los dos años anteriores (ver Figura 2.3).

Además del *IF*, el *JCR* utiliza el *Five-Year Impact Factor* (Jacso, 2009) el cual ofrece una ventana mayor de tiempo (5 años) para el recuento de las citas. Las últimas incorporaciones al *JCR* son el *Eigen factor* y el *Article influence score*. El *Eigen factor*(Bergstrom, 2007) fue desarrollado por Bergstrom, en la Universidad de Washington. Su nombre proviene del *eigen vector*, una medida de centralidad general empleada en el análisis de redes que se basa en las conexiones que un nodo mantiene con otros nodos de gran importancia y que se calcula

[10] http://thomsonreuters.com/products_services/science/science_products/a-z/journal_citation_reports/#tab2

Figura 2.3: Figura tomada de Thomsonreuters.com, para el cálculo del IF

como una suma ponderada de los caminos que se originan en un nodo. Calcula las citas recibidas en los últimos 5 años, normaliza el prestigio de la revista basándose en el número de las referencias citadas dentro de la ventana de la publicación (5 años) y elimina las auto-citas que realiza la revista(Colledge et al., 2010; Torres-Salinas and Jiménez-Contreras, 2010).

El *Eigen factor* clasifica las revistas de una manera similar a la utilizada por Google en su algoritmo PageRank para clasificar la importancia de los sitios Web en una búsqueda(Fersht, 2009). El *Article influence score (AIS)* se deriva del *Eigen factor* y es conceptualmente similar al *IF* en que existe un numerador y un denominador (el número de documentos citables) excepto que utiliza al *Eigen factor* (en lugar del número total de citas) como numerador. Con lo cual, difiere de *IF* donde todas las citas son contabilizadas por igual independientemente de su origen, en cambio en *AIS*, cada cita es multiplicada por la "calidad" de las revistas que citan, lo que resulta en un mayor peso de las citas que provienen de revistas altamente citados, y menor peso de las revistas con bajo número de citas.

Para facilitar la interpretación, el *AIS* se normaliza, por lo que el artículo promedio en *JCR* cuenta con *AIS* de 1,00(Rizkallah and Sin, 2010). Una puntuación superior a 1,00 indica que los artículos en esa revista tienen una influencia superior a la media, una puntuación inferior a 1,00 indica que los artículos en la revista tienen una influencia inferior a la media [11].

[11] http://admin-apps.webofknowledge.com/JCR/help/h_eigenfact.htm

2.3.3. SCImago Journal and Country Rank

Por su parte, *Scopus* posee el *SCImago Journal and Country Rank (SJR)* desarrollado por *SCImago Research Group*, con sede en España y dirigido por el Profesor Félix de Moya y el *Source Normalized Impact per Paper (SNIP)* desarrollado por el Centre for Science and Technology Studies (CWTS), con base en Leiden, Netherlands, y dirigido por el Profesor Anthony van Raan. El SJR utiliza como fuente de datos las revistas indexadas por Scopus desde 1996, este indicador pertenece a la nueva familia de indicadores basados en la *eigenvector centrality*. El *SJR* es una métrica de tamaño independiente destinado a medir el actual "prestigio promedio por artículo" de las revistas utilizadas en los procesos de evaluación de la investigación(González-Pereira et al., 2009).

Una de las limitaciones del análisis tradicional de citas es que todas las citas se consideran "iguales", es decir, que posee el mismo peso tanto una cita de una revista multidisciplinar y muy leída como de una de interés local. *SJR* es un indicador de prestigio inspirado por el algoritmo *PageRank* (Page et al., 1998) de *Google*, por lo que el campo de interés de la revista, la calidad y la reputación tienen un efecto directo sobre el valor de las citas que da a otras revistas.

La idea básica es que cuando una revista A es citada, por ejemplo, 100 veces por las revistas de más alto rango en el campo, recibe más prestigio que una revista B que también recibe 100 citas, pero de revistas de menor prestigio con una baja visibilidad. *SJR* hace una distinción entre la *popularidad de una revista* y el *prestigio de una revista*. Se podría decir que las revistas A y B tienen la misma popularidad, pero A tiene un prestigio mayor que B. A y B tienen el mismo *JIF*, pero A tendría un mayor *SJR* que B.

Generalizando, *JIF* se puede considerar como una medida de la popularidad, ya que resume todas las citas que recibe una revista, independientemente de la situación de las revistas que citan, mientras que SJR mide el prestigio. La idea de la recursividad, o cálculo iterativo, es esencial. Paso a paso, *SJR* asigna los pesos a las citas en el paso actual de acuerdo al *SJR* de la revista que está citando en el paso anterior. Bajo ciertas condiciones, este proceso converge para que los valores *SJR* no cambien significativamente con pasos adicionales, y al final una cita de una fuente con un relativamente alto *SJR* vale más que una cita de una fuente con un relativamente bajo *SJR*.

SJR también tiene como objetivo limitar los beneficios excesivos derivados de las autocitas de la revista. En el cálculo de *SJR*, se descuentan las autocitas una vez que se supera un tercio del total de las citas que recibe una revista. De esta manera, el valor de las autocitas de la revista sigue siendo reconocido, pero *SJR* intenta limitar lo que a menudo se considera las prácticas de manipulación de citas (Colledge et al., 2010).

En lo que respecta a *SNIP*, este indicador ha sido propuesto por (Moed, 2009). Se basa en la idea de la "Citation potential" (citación potencial) introducida por Garfield (Garfield, 1979) y que se entiende como el número de referencias medio que contienen los artículos de un área determinada. La citación potencial en un campo de una revista, es una medida de las características de citación en la que se encuentra el campo de la misma, determinado por la frecuencia y la rapidez con que los autores citan otras obras, y lo bien que su campo está cubierto por la base de datos en cuestión.

La citación potencial puede ser concebida como una medida de la actualidad del campo. Los campos que se actualizan constantemente tienden a atraer a muchos autores que comparten un interés intelectual, y en este sentido se puede calificar como "popular". Los artículos son escritos en un número limitado de publicaciones periódicas de gran visibilidad, y los autores suelen citar, aparte de la base intelectual común, los trabajos más recientes de sus colegas. Estos campos populares tienden a tener mayores *JIF*.

El campo temático de una revista se define como el conjunto de artículos que citan la revista. De esta forma se define el indicador *SNIP* como el número de citas medio recibido por los artículos de una revista durante tres años, que denomina *Raw Impact per Paper (RIP)* el cual es muy similar al *JIF*, dividido entre la citación potencial del campo científico de la revista que sintetiza en el indicador *Relative database citation potential(RDCP)*. *SNIP* tiene en cuenta no solo las diferencias entre, sino también dentro de las categorías temáticas de la revistas.(Torres-Salinas and Jiménez-Contreras, 2010; Colledge et al., 2010).

2.3.4. Google Scholar Citacions

Google Scholar Citacions (GSC) [12] también llamado *Google Scholar Author Citation Tracker (GSACT)* fue oficialmente lanzado en pruebas en julio de 2011, como una forma sencilla de que los científicos y académicos puedan realizar un seguimiento de las citas de sus artículos. Se puede revisar quien cita las publicaciones y graficar las citas a través del tiempo. *GSACT* es un módulo de *Google Scholar*, y representa un conjunto pequeño de la base de datos completa de *Google Scholar*.

Los autores se dan de alta con una cuenta de *Google*, crean un perfil que consiste en datos personales (nombre, afiliación, disciplinas de interés) y una vez que la cuenta de email es verificada, puede editar los registros de *Google Scholar*, corrigiendo (normalizando) la información del buscador, unir registros duplicados, e incluso añadir de forma manual otros trabajos que hayan sido indexados por *Google*.

Además presenta tres indicadores bibliométricos: el número total de citas de los trabajos, el h-index del investigador, y el $i10$-index, esto es el número de trabajos con más de diez citas, tanto para toda la carrera académica como para el periodo más reciente. Pero quizá lo más interesante es que la información sobre las citas recibidas y la producción se actualizan de forma automática a medida que va siendo indexada por *Google*, sin necesidad de concurso por parte del académico, que encuentra siempre su información al día(Cabezas-Clavijo and Torres-Salinas, 2012).

Este nuevo módulo ofrece algunas opciones como ordenar la lista de los resultados por año de publicación, título y citas recibidas. Como lo indica (Jacso, 2012), las pruebas preliminares arrojan resultados alentadores, principalmente por la limpieza de los registros en las muestras testeadas, y por la participación de los usuarios registrados que pueden corregir los errores encontrados en sus perfiles académicos. Claramente el problema de la desambiguación, eliminación o fusión de registros duplicados, eliminación de publicaciones no pertenecientes al autor es una tarea delegada al autor, con lo cual el problema sigue sin ser resuelto en su totalidad.

Algunos autores como Anne-Wil Harzing, expresan sus temores acerca de la honestidad de los autores, y si estos no serán capaces de falsear

[12] `https://www.scholar.google.com/citations`

sus datos en busca de un ascenso académico o como manera de aumentar su visibilidad en la web. Cualquiera puede falsear sus datos para aparecer en posiciones prominentes, sin embargo parece poco probable ya que sería rápidamente detectado por el resto de usuarios, cayendo en el descrédito científico (Herther, 2011).

Para el éxito de esta herramienta se necesitará por un lado que todos los científicos creen un perfil, sin importar la popularidad de estos, ya que es probable que científicos que posean un pequeño número de publicaciones no se atrevan a crear un perfil público donde se haga evidente esto, y en el otro extremo, científicos muy productivos estarían alentados a hacerlo, además es deseable que los autores dediquen algo de tiempo para realizar las correcciones necesarias, y por otro lado será necesario confiar en ellos, confiar en que lo que está publicado no solo pertenece a dicho autor, sino que es contenido real y no infringe las reglas o no representa contenido malicioso.

2.3.5. Google Scholar Metrics

La aparición de *Google Scholar Metrics* (GSM) [13] en abril de 2012 como nuevo sistema de evaluación bibliométrica de revistas científicas a partir del recuento de las citas bibliográficas que éstas han recibido en *Google Scholar* (recuento de citas en revistas, actas, repositorios) rompe el duopolio ejercido hasta el momento por las bases de datos *Web of Science* y *Scopus* (Delgado López-Cózar and Caballero, 2013).

Las revistas, actas y repositorios incluidos en este índice, deben haber publicado al menos 100 artículos y haber recibido al menos una cita en los últimos cinco años, de otra manera no pueden ser incluidos. Este indicador se presenta como una alternativa a los rankings tradicionales (JCR y SJR), principalmente por la popularidad y la gran aceptación que *Google Scholar* está recibiendo en los últimos años.

GSM posee estadísticas para publicaciones en 12 lenguajes a saber: inglés, chino, portugués, español, alemán, ruso, francés, japonés, coreano, polaco, ucraniano, indonesio. Se pueden consultar las publicaciones más destacadas según áreas y disciplinas, pero esta clasificación se encuentra disponible solo para el idioma inglés, las áreas principales son: Physics & Mathematics, Chemical & Material Sciences, Engineering & Compu-

[13] https://scholar.google.com/citations?view_op=top_venues

ter Science, Health & Medical Sciences, Life Sciences & Earth Sciences, Humanities, Literature & Arts, Business, Economics & Management y Social Sciences (Google, 2016).

En GSM, los rankings de publicaciones se clasifican por impacto (h-index) y se pueden consultar por idiomas, cada una de las cuales muestra las 100 publicaciones (revistas, actas, repositorios) de mayor impacto (Martín-Martín et al., 2016b). En la Figura 2.4 se puede observar el Top 20 de publicaciones en inglés según Google Scholar Metrics 2016. En (Martín-Martín et al., 2016a) se presenta una revisión muy actualizada de las mejoras, errores, correcciones, sustracción, agregados y cambios en la última versión (la 2016) de este indicador.

Publicaciones principales - inglés Más información

Publicación	Índice h5	Mediana h5
1. Nature	379	560
2. The New England Journal of Medicine	342	548
3. Science	312	464
4. The Lancet	259	418
5. Cell	224	339
6. Chemical Society reviews	224	329
7. Journal of the American Chemical Society	218	293
8. Proceedings of the National Academy of Sciences	215	286
9. Advanced Materials	201	301
10. Angewandte Chemie International Edition	198	276
11. Journal of Clinical Oncology	197	265
12. Physical Review Letters	196	282
13. Chemical Reviews	194	332
14. Nano Letters	192	270
15. JAMA	189	269
16. Nucleic Acids Research	184	345
17. Energy & Environmental Science	184	254
18. ACS Nano	180	243
19. Nature Genetics	179	267
20. arXiv Cosmology and Extragalactic Astrophysics (astro-ph.CO)	176	243

Figura 2.4: Top 20 de publicaciones según Google Scholar Metrics 2016

Como todos los productos derivados de *Google Scholar*, este indicador no está exento de poseer errores, omisiones o de ser susceptible al fraude como lo indican (Delgado López-Cózar et al., 2012) en su experimento, en el cual demuestran que tan sencillo es alterar los índices y recuento de

citas de trabajos asociados a un investigador o grupo de investigadores. De un modo u otro, GSM se puede presentar como alternativa válida a los índices bibliométricos tradicionales por todas las bondades que ya se conocen de su fuente de datos, *Google Scholar*.

2.3.6. Indicadores de la productividad individual

Los indicadores nombrados anteriormente se encargan de evaluar y comparar revistas, otra clase de indicadores bibliométricos son utilizados para evaluar la productividad de un científico o también son utilizados para comparar dos o más científicos en un mismo campo (vale aclarar que al comparar el valor de un indicador de productividad entre científicos, estos deben ser del mismo campo ya que existen campos donde se publica muy poco o donde es más complicado tener un número importante de publicaciones).

Se debe tener en cuenta que los indicadores para evaluar una revista no pueden y no deben ser utilizados para evaluar un científico, como lo indica (Van Noorden, 2010) "Si hay algo en lo que todos los estudiosos de las variables bibliométricas están de acuerdo, es que nunca se debería usar el *IF* de una revista para medir el impacto de un artículo o de un científico. Eso es un pecado mortal". El *IF* es de poca utilidad para medir el desempeño de un individuo, sino que se aplica solo a la popularidad de una revista.

El indicador pionero en este tema es el h-index (Hirsch, 2005), un índice fácilmente calculable, que presenta una estimación de la importancia, el significado y el impacto general de las contribuciones de investigación acumuladas para un científico. Según Hirsch, el creador de este indicador, lo define como: "Un científico tiene índice h si h de sus N_p trabajos (publicaciones) tienen al menos h citaciones cada uno, y los otros $(N_p - h)$ trabajos tienen menos que h citaciones cada uno". Como se observa, este indicador es muy sencillo de calcular y tiene en cuenta tanto la cantidad como el impacto de las publicaciones.

Según (Schreiber, 2008a), el h-index es robusto en el sentido que no es sensible a los artículos no citados o a aquellos que reciben muy pocas citas, esto claramente es una ventaja comparado con indicadores como el número medio de citas por artículo o el número total de artículos. Sin embargo, el h-index también es robusto en el sentido que no es sensible a

uno o varios artículos extraordinariamente muy citados, porque una vez que una publicación ha alcanzado el conjunto-h definitorio (h-defining set) o también llamado h-core (que son los h artículos más citados del autor), no se vuelven más relevantes si reciben o no futuras citas (esta es la principal desventaja del h-index).

Algunos autores como (Bornmann et al., 2008; Bar-Ilan, 2008) se han encargado de señalar esta y otras desventajas del uso de este indicador, por ello se han propuesto un interesante número de variantes del mismo. El f-index y t-index, ambos propuestos por (Tol, 2009), utilizan la media armónica y media geométrica respectivamente, para el cálculo de estos; el A-index (Jin, 2006) Este índice se define simplemente como el número medio de citas recibidas por las publicaciones incluidas en el h-core. El nombre de este índice se deriva del hecho de que es solo un promedio (A del inglés *average*). Matemáticamente:

$$A = \frac{1}{h} \sum_{j=1}^{h} cit_j$$

En esta fórmula, cit_j representa el número de citas. El A-index utiliza los mismos datos que el h-index de manera que el problema de la precisión es exactamente el mismo que para el h-index. Este índice posee una desventaja para valores altos de h, ya que penaliza a los autores con h-index grande.

El R-index (Jin et al., 2007), introduce una mejora sobre el A-index, la fórmula matemática es la siguiente (R del inglés *root*):

$$R = \sqrt{\sum_{j=1}^{h} cit_j}$$

En el caso especial de que cada cit_j sea exactamente igual a $h, R = h$, esta es una ventaja de utilizar la raiz cuadrada de la suma y no la suma en misma.

Debido a que el h-index no puede disminuir y que los científicos pueden, por así decirlo, "dormirse en sus laureles", (Jin, 2007) propuso la siguiente adaptación del R-index, denominada AR-index por ser un indicador que depende de la edad (A del inglés *age*) de las publicaciones:

$$AR = \sqrt{\sum_{j=1}^{h} \frac{cit_j}{a_j}}$$

Donde a_j indica la edad del artículo j. Si todos los cit_j son iguales a h y todos los a_j son iguales a uno, entonces $AR = h$. Esta es una buena idea ya que, claramente, para fines de evaluación de la investigación, el trabajo realizado hace veinte años resulta de menor importancia que el trabajo realizado hace cuatro años (Rousseau, 2008). Para (Jin, 2007), el par (h, AR) representa un indicador significativo para la evaluación de la investigación.

Una de las variantes del h-index que ha tenido bastante atención es el llamado g-index (Egghe, 2006b,a), este indicador fue diseñado para dar mayor importancia a los artículos más citados del autor, es decir, el g-index caracteriza mejor el conjunto de datos que el h-index. El g-index se define como: "Un conjunto de artículos tienen g-index g si g es el rango más grande tal que el top g de artículos, en conjunto, tienen al menos g^2 citaciones. Esto también quiere decir que el top $g+1$ de artículos tiene menos que $(g+1)^2$ citas". Esta definición es equivalente a la determinación del g-index como el mayor número de artículos que recibieron en promedio g o más citas (Schreiber, 2010b).

Sin embargo, aunque el g-index resulta adecuado en la evaluación de la producción de un investigador porque incorpora el número real de citas que reciben sus publicaciones, presenta el inconveniente que se ve fuertemente influenciado por artículos muy exitosos, es decir, artículos que poseen un gran número de citas.

Tanto el h-index como el g-index miden aspectos diferentes del conjunto de publicaciones de un autor. Tomados en conjunto, g y h presentan un cuadro conciso de los logros de un científico en términos de publicaciones y citaciones (Rousseau, 2006). En este sentido (Alonso et al., 2010) propone un índice combinado, llamado hg-index que trata de fusionar todas las ventajas de las dos medidas anteriores e intenta minimizar los inconvenientes que cada uno de ellos presentan. Se define como: "El hg-index de un investigador es calculado como la media geométrica de sus h-index y g-index". La fórmula es la siguiente:

$$hg = \sqrt{h.g}$$

Este nuevo indicador es sencillo de obtener una vez calculados los h-index y g-index. Al tener en cuenta ambas dimensiones (número de publicaciones y cantidad de citas) proporciona una manera más fina de comparar a dos o más científicos. Además, al resolver las principales desventajas de ambos indicadores (el h-index no es sensible a los artículos muy citados y el g-index es sensible a un único artículo muy citado) logrando así un mejor equilibrio entre el impacto de la mayor parte de los mejores trabajos del autor y los muy citados.

Creado por *Google Scholar* y utilizado en la sección *Google Scholar Citation*, i10-*Index* es el número de publicaciones con al menos 10 citas. Esta medida muy simple es utilizada solamente por *Google Scholar*, y es otra manera de ayudar a medir la productividad de un científico Cornell University Library (2017).

Para un mayor detalle de estos indicadores revisar los trabajos de (Schreiber, 2010a; Alonso et al., 2010).

Si bien el h-index no puede ser utilizado para comparar académicos que trabajan en diferentes disciplinas o se encuentran en etapas diferentes de su carrera, (Harzing et al., 2014) introdujo hace poco una métrica llamada *hI,annual* o *hIa*. El hIa-index representa el incremento promedio anual en el h-index individual, esta métrica atenúa las diferencias en el h-index atribuibles a cuestiones de fondo de cada especialidad (como la coautoría) y al largo de la carrera académica. El *hIa*-index es calculado dividiendo el h-index individual (un h-index corregido por el número de coautores) por el número de años que se encuentra activo un académico, es decir, el número de años que han transcurrido desde su primera publicación (Harzing and Mijnhardt, 2015).

Si bien los indicadores que son utilizados para evaluar una revista no pueden ser utilizados para evaluar un científico, como es el caso del *IF*, en el sentido contrario puede resultar bastante útil, es decir, utilizar indicadores para evaluar la productividad de un científico para evaluar una revista. En este sentido el *h-index* ha dado buenos resultados al ser utilizado como medida del rendimiento de revistas, grupos de investigación, departamentos y países(Thor and Bornmann, 2011; Harzing and van der Wal, 2009).

Utilizando los datos de las citas como elemento para evaluar la investigación científica, en última instancia significa utilizar las estadísticas del análisis basado en citas para clasificar cosas tales como revistas, docu-

mentos, personas, programas y disciplinas. Las herramientas estadísticas utilizadas para clasificar estas cosas son a menudo mal interpretadas y mal utilizadas (Adler et al., 2009), por ejemplo:

- Para las revistas, el factor de impacto se utiliza con mayor frecuencia para la clasificación de estas. Se trata de un promedio simple derivado de la distribución de citas para una colección de artículos de la revista. Este promedio únicamente captura una pequeña cantidad de información acerca de la distribución, es una estadística aproximada. Además, existen muchos factores de interferencia al juzgar revistas por medio de las citas, y cualquier comparación entre revistas requiere precaución al utilizar el factor de impacto. La sola utilización del factor de impacto para juzgar una revista es como utilizar el peso de una persona para juzgar su salud.

- Para los artículos, en lugar de valerse del recuento de citas para comparar artículos individuales, las personas sustituyen con frecuencia el factor de impacto de las revistas en las que aparecen los artículos. Estas personas creen que un factor de impacto alto debe indicar un gran número de citas. Pero a menudo no es el caso. Este es un mal uso generalizado de las estadísticas que necesita ser evitado en cualquier momento y dondequiera que ocurra.

- Para los científicos individuales, los registros completos de citas pueden ser difíciles de comparar. Como consecuencia, ha habido intentos de encontrar estadísticas simples que capten la complejidad de un registro de citas para un científico, con un único número. El más notable de ellos es el h-index, que parece ser cada vez más popular. Pero incluso una inspección al azar del h-index y sus variantes, demuestra que estos son intentos sencillos por entender los complejos registros de citas. Mientras que estos indicadores capturan una pequeña cantidad de información acerca de la distribución de las citas de un científico, descartan información crucial que es esencial para evaluar la investigación.

2.3.7. Problemas comunes

En los motores de búsqueda académicos no solo la cobertura representa un problema. Como lo indica (Torres-Salinas et al., 2009), la amplia

cobertura, la variedad de fuentes de información empleadas y el procesamiento automático de la información llevan consigo la ausencia de normalización en los datos de *Google Scholar*. Esto redunda en el empleo de mayores controles y por supuesto en el incremento de la cantidad de tiempo para obtener resultados que satisfagan las búsquedas realizadas. Estas bases de datos no poseen muchos controles de consistencia, muchas veces debido a que son bases de datos de gran tamaño o porque el motor de búsqueda/indexado es similar a los clásicos motores de búsqueda web, como es el caso de *Google Scholar* el cual proviene de *Google Web Search* (Beel and Gipp, 2010).

Los estudios de (Meho and Yang, 2006), para realizar rankings bibliométricos, revelan lo diferente que resulta procesar los datos provenientes de *Wos*, *Scopus* y GS. En dicho análisis, se indica que se tardó más tiempo en recoger y analizar los datos de *Google Scholar* para dos profesores que lo que tomó recoger y analizar los datos de los 22 profesores de la *Web of Science* y *Scopus* en conjunto. De igual forma, en (Meho and Yang, 2007), las tareas de recolectar, depurar, estandarizar y cargar los datos en las herramientas de análisis, demandó al rededor de 100 horas para *WoS*, 200 horas para *Scopus* y casi 3000 horas para GS.

Esta es una de las desventajas de GS, ya que los resultados son recuperados de un modo que resulta poco práctico para grandes cantidades de datos o un número grande de participantes dentro del estudio. Diferente a *WoS* y *Scopus*, GS no permite ordenar los resultados en ninguna forma (como por fecha, nombre de autor u origen de datos), los resultados son clasificados según cuán relevantes son a la consulta realizada. Otra desventaja de GS incluye las citas duplicadas (por ejemplo, contabilizando una cita publicada en dos diferentes formas, como pre-impreso y artículo de revista como dos citas) (Meho and Yang, 2006)

Es evidente que *Google Scholar*, al incluir indiscriminadamente todas las citas que es capaz de identificar en cualquier documento, no puede asegurar ningún control de calidad de la información científica que presenta. Esta es la diferencia entre un entorno controlado (bases de datos tradicionales) y uno no controlado. Esta situación pone en evidencia que, por el momento, su utilización a media y gran escala como herramienta de evaluación científica supone un consumo de recursos tan grande que la inhabilita.

La falta de normalización es un problema común a la mayoría de las

fuentes de datos y un tema recurrente en la mayoría de los estudios basados en el análisis de citas (Waltman, 2016; Vanclay, 2012; Miguel and Solana, 2010). Entiéndase por falta de normalización registros duplicados, registros con errores ortotipográficos tratados como registros diferentes, trabajos atribuidos a otros autores, resultados imprecisos. Claro está que las bases de datos bibliográficas tradicionales poseen numerosos controles y realizan un gran esfuerzo para minimizar este problema, pero esto no quiere decir que dichas bases de datos se encuentren exentas de este problema, como lo indica (Valderrama-Zurián et al., 2015) quien encontró que la existencia de registros duplicados en *Scopus* se debía principalmente a cambios en el nombre de la revista indexada, diferencias ortográficas en los títulos de revistas y variaciones de títulos de revistas.

Mientras que *Google Scholar* poco a poco va incrementando la calidad de los resultados, aun posee serios inconvenientes si no es tratado con cuidado y con las precauciones necesarias. Aunque existe literatura al respecto encargada de documentar esta falta de normalización, los errores, fallas y problemas técnicos de esta herramienta(Aguillo, 2012; Jacso, 2008a,b, 2005), *Google Scholar* no deja de ser el objeto de análisis y comparación en numerosos estudios, por ello en este trabajo se presentará un mecanismo para resolver los problemas más comunes a la hora de realizar una búsqueda en un intento por ajustar lo mejor posible los resultados de *Google Scholar* a la realidad.

A diferencia del motor de búsqueda genérico *Google* y otros motores de búsqueda, cuando muestra los resultados de la búsqueda, *Google Scholar* utiliza *de-duplication*, un proceso para remover los registros duplicados, sin embargo este proceso no asegura que el registro que se mantiene sea el mejor de una publicación dada. La calidad en la selección de los registros es ciertamente una área donde *Google Scholar* necesita realizar mejoras (Chen, 2010).

2.3.8. Web spam

La Recuperación de Información Acusatoria (Fetterly, 2007) estudia la forma de realizar las tareas de recuperación de información, tales como la búsqueda o la clasificación, en las colecciones en las que algunos objetos han sido manipulados maliciosamente. La forma más frecuente de tal manipulación es el *spam*, un problema que prevalece en la mayoría de las

comunicaciones electrónicas.

El Web *spam* se manifiesta como contenido web generado deliberadamente con el propósito de desencadenar relevancia injustificadamente favorable o importancia a una página o conjunto de páginas web (Gyongyi and Molina, 2005). El *spam* ha sido identificado como uno de los principales desafíos que los motores de búsquedas basados en la web deben abordar (Henzinger et al., 2002), ya que no solo deteriora la calidad de los resultados de búsqueda, sino que también debilita la confianza entre el usuario y el proveedor de motores de búsqueda, y desperdicia una cantidad significativa de recursos computacionales en el motor de búsqueda (Abernethy et al., 2010).

Google Scholar al igual que otros motores de búsqueda académicos no están libres del SPAM de citas, algunos autores por elevar el número de citas de sus trabajos o incrementar los indicadores de productividad, publican contenido no oficial, publican bajo el dominio web de universidades o simplemente publican en varios sitios el mismo contenido. En (Beel and Gipp, 2010) se puede ver un análisis detallado de que *Google Scholar* no está libre de SPAM y de cuales son los mecanismos más utilizados, por otro lado, en dicho trabajo se generan artículos con contenido aleatorio utilizando la herramienta *SciGen* [14] para luego permitir que sean indexados por los rastreadores de *Google Scholar*.

Del mismo modo (Delgado López-Cózar et al., 2012) realiza un experimento para manipular los indicadores de productividad de un equipo de investigadores mediante trabajos de un autor ficticio que cita a los trabajos de este equipo. En un trabajo posterior de (Delgado López-Cózar et al., 2014), para analizar la capacidad de *Google Scholar* de detectar la manipulación del recuento de citas, para ello crearon 6 documentos que fueron subidos a una web institucional, estos trabajos pertenecían a un autor ficticio y referenciaban a todos los trabajos de los miembros del grupo de investigación EC3 [15] de la Universidad de Granada. *Google Scholar* al detectar estos artículos disparó el número de citas en los perfiles de GSC de los autores, en la Figura 2.5 (imagen tomada de (Delgado López-Cózar et al., 2014)) se muestran los resultados de este pequeño experimento.

La presión por publicar avivada por los sistemas de evaluación del

[14] http://pdos.csail.mit.edu/scigen/
[15] https://ec3metrics.com/

		Bibliometric Indicators								
		Nr Citations			H-Index			I10-Index		
Author Research Profile	*Time Period*	*Before and After manipulation*			*Before and After manipulation*			*Before and After manipulation*		
Emilio Delgado López-Cózar	**All years**	862	→	1297	15	→	17	20	→	40
	Since 2007	560	→	995	10	→	15	11	→	33
Nicolás Robinson-Garcia	**All years**	4	→	29	1	→	4	0	→	0
	Since 2007	4	→	29	1	→	4	0	→	0
Daniel Torres-Salinas	**All years**	227	→	416	9	→	11	7	→	17
	Since 2007	226	→	415	9	→	11	7	→	17

Figura 2.5: Valores de h-index y i10-index de acuerdo a GSC antes y después del experimento

rendimiento científico adoptados en todos los ámbistos y paises puede exacerbar las malas artes y prácticas comunicativas de los científicos a fin de manipular la orientación y el sentido de los números. Y es que, a día de hoy no existen más controles o filtros para evitar el fraude en los datos que las reservas éticas y morales de los propios investigadores (Cabezas-Clavijo and Delgado-López-Cózar, 2012).

2.4. Recuperación de información, herramientas

Con los avances y mejoras implementados en los motores de búsqueda, y el aumento de datos públicos en Internet, el usuario debe hacer frente a un gran reto, encontrar información y conocimiento útiles. Existe una constante demanda de motores de búsqueda con más inteligencia y nuevas funcionalidades para apoyar las búsquedas(Zeng et al., 2009). La creación de *Google Scholar*, un motor de búsqueda académico dedicado a rastrear la Web en busca de literatura científica, revolucionó el mundo de los sistemas de búsqueda de información científica. Particularmente, tuvo un efecto muy positivo en aquellas disciplinas donde los hábitos de publicación no estaban limitados a publicar en revistas científicas (Martín-Martín et al., 2016b).

Desde noviembre del 2004, *Google Scholar* ha sido considerado una posible alternativa a *ISI WoS* y *Scopus*. En primer lugar por ser una herramienta de libre acceso, lo cual provee una mayor transparencia y permite que cualquier persona pueda realizar un análisis de citas (Pauly and Stergiou, 2005). Y en segundo lugar porque proporciona un me-

dio para buscar bibliografía especializada a través de muchas disciplinas y fuentes: artículos revisados por pares, tesis, libros, resúmenes y artículos provenientes de editoriales académicas, sociedades profesionales, repositorios digitales, universidades y otras organizaciones académicas (Baneyx, 2008).

Estudios recientes evidencian que los estudiantes tienden a preferir *Google Scholar* sobre bases de datos bibliográficas convencionales debido a su simplicidad, su velocidad y su similitud con *Google* (Dixon et al., 2010; Orduña-Malea et al., 2016a). Por todo esto *Google Scholar* se ha convertido en uno de los motores de búsqueda académicos más populares y controvertidos, esto último debido al bajo control de la calidad de la información que indexa. En los resultados de las búsquedas no todo es documentación oficial, entre los resultados no faltan las presentaciones *PowerPoint*, resúmenes y programas de asignaturas, contenido no académico, entre tantos.

2.4.1. Publish or Perish

El análisis de citas ofrece un medio para cuantificar el impacto del trabajo de los científicos, realizar este análisis basado en motores web de libre acceso, como es el caso de *Google Scholar* y recientemente MAS y MA (*Microsoft Academic*, que es la versión 2 de MAS), es una tarea que requiere una considerable cantidad de tiempo, como se ha indicado anteriormente, debido a la enorme cantidad de información y a la diversidad de contenido que esta herramienta ofrece, el cálculo de indicadores fiables no es sencillo. En este sentido, para automatizar parte de esta tarea, existe la herramienta conocida como *Publish or Perish* (PoP) (Harzing, 2010), un software desarrollado por Anne-Wil Harzing, profesora en la Universidad de Melbourne, Australia.

La principal utilidad de PoP es listar los resultados de *Google Scholar* y exportarlos. Sin embargo, esta herramienta posee dos limitaciones. La primera de ellas es que no permite fusionar el número de citas cuando un artículo aparece varias veces en *Google Scholar*, lo cual es muy frecuente. En segundo lugar, se puede buscar los artículos de Audrey Baneyx (término de búsqueda "A Baneyx") y obtener, por ejemplo, un resultado de Francois Baneyx porque este autor ha publicado con A Bianchi. Del mismo modo, también se puede obtener un resultado para Alexandra Baneyx porque la búsqueda "A Baneyx" es demasiado grande, pero

la búsqueda "Audrey Baneyx" es demasiado limitado. El uso de PoP requiere depurar la lista de las publicaciones obtenidas y calcular los indicadores de los datos similares de nuevo (Harzing, 2007). Además de algunas simples estadísticas (número de documentos, número de citas, y otros), PoP calcula las siguientes métricas (Harzing, 2016b):

1. h-index.
2. g-index.
3. e-index (Zhang, 2009). El e-index es la (raíz cuadrada) de los excedentes de citas ignorados en el h-set más allá de h^2, es decir, más allá del mínimo teórico requerido para obtener un h-index de "h". El objetivo del e-index es diferenciar entre los científicos con similares h-index, pero con diferentes patrones de citas.
4. *Contemporary* h-index (Sidiropoulos et al., 2007). Su objetivo es mejorar el h-index, dando mayor peso a los artículos recientes, premiando de este modo a los científicos que mantienen un nivel constante de actividad.
5. AR-index.
6. *Individual* h-index (Batista et al., 2006). Se divide el h-index estándar por el número promedio de autores en los artículos que contribuyeron al h-index, con el fin de reducir los efectos de la co-autoría.
7. *Individual* h-index (variante PoP). PoP también implementa una alternativa del *individual* h-index que toma un enfoque diferente: en lugar de dividir el total h-index, en primer lugar, se normaliza el número de citas para cada trabajo dividiendo el número de citas por el número de autores de ese documento, a continuación, se calcula el h-index de los recuentos de citas normalizadas.
8. *Multi-authored* h-index, conocido como h_m-index (Schreiber, 2008b). Este método utiliza recuentos fraccionados de publicaciones en lugar de reducciones del número de citas para el recuento de la autoría compartida de las publicaciones, y luego determina el *multi-authored* h_m-index basado en el rango efectivo resultante de los trabajos utilizando los recuentos de citas concentrados.
9. *hI,annual*

2.4.2. Scholarometer

Otra de las herramientas que utiliza como fuente de datos a *Google Scholar* es *Scholarometer*[16]. *Scholarometer* es una herramienta social para facilitar el análisis de citas y ayudar a evaluar el impacto de las publicaciones de un autor. Fue desarrollada en la Universidad de Indiana y lanzada en Noviembre de 2009, con el doble objetivo de explorar el enfoque *crowdsourcing* para anotaciones disciplinarias y métricas de impacto interdisciplinarias (Hoang et al., 2010). Estos dos objetivos están íntimamente relacionados y se refuerzan mutuamente.

Las anotaciones permiten la recopilación de estadísticas específicas de la disciplina, y por lo tanto el cálculo de métricas de impacto universal. Por su parte, el servicio prestado a los usuarios mediante el cálculo de estas métricas funciona como un incentivo para que los usuarios den las anotaciones. El *crowdsourcing* es un enfoque para aprovechar los conocimientos de una comunidad a través de plataformas web con el fin de resolver problemas prácticos. *Scholarometer* aplica *crowdsourcing* para anotaciones académicas. Los usuarios proporcionan anotaciones disciplinarias a cambio del acceso a los datos de las citas obtenidas, provenientes de la consulta a los servicios bibliográficos. Con lo cual, se consideran dos fuentes de datos: (i) datos de citas bibliográficas on-line desde una biblioteca digital (GS) y (ii) anotaciones suministradas por los usuarios sobre los autores con las etiquetas de la disciplina consultada(Sun et al., 2013).

La herramienta *Scholarometer* tiene dos interfaces para la comunicación con los usuarios: una en la extensión del navegador para ingresar las consultas y las etiquetas, y la otra en la ventana principal del navegador para la presentación y manipulación de los datos bibliográficos y resultados del análisis de citas. La extensión del navegador está disponible en dos versiones: uno para el navegador *Firefox* alojado en el sitio *Mozilla Firefox Add-ons*, y otro para el navegador *Chrome* alojado en el sitio de *Google Chrome Extensions*.

La arquitectura y flujo de trabajo de *Scholarometer* se ilustra en la Figura 2.6. Consta de 6 pasos: (1) En primer lugar, el usuario introduce una consulta y etiquetas de disciplina para un autor en un formulario de búsqueda proporcionado por la extensión del navegador. (2) La ex-

[16] `http://scholarometer.indiana.edu`

tensión del navegador reenvía la consulta a *Google Scholar*. (3) *Google Scholar* devuelve los resultados de la consulta a la extensión del navegador. (4) Luego, la extensión del navegador envía los resultados al servidor *Scholarometer*. Este analiza los resultados para extraer las citas y otros metadatos, que luego se insertarán en la base de datos, junto con los metadatos de las anotaciones. (5) El servidor de *Scholarometer* envía al navegador del cliente los registros bibliográficos y las métricas de impacto para el autor(es) consultados. (6) Por último, el navegador del cliente reproduce los datos de manera interactiva. El usuario visualiza los resultados en una nueva pestaña del navegador y puede realizar acciones avanzadas como la clasificación, filtrado, eliminación y fusión de registros(Hoang et al., 2010).

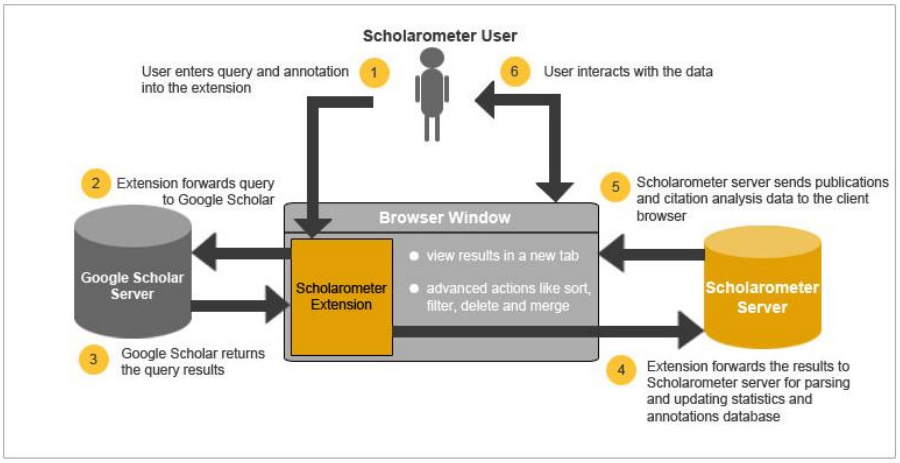


Figura 2.6: Arquitectura y flujo de trabajo de Scholarometer

Para realizar una consulta, los usuarios deben ingresar por un lado el nombre del autor y por otro las etiquetas (anotaciones) de la disciplina asociada a dicho autor. Los datos que se recogen provienen de los usuarios, resulta natural que posean ruido, con lo cual se emplea una lista negra (*blacklist*) para prevenir que los *spammers* corrompan la base de datos, ejemplo de ello es el autor ficticio "Ike Antkare", creado para resaltar la vulnerabilidad de las bases de datos de citas *on-line* (Labbé, 2010). Cuando una consulta coincide con un nombre en la *blacklist*, el sistema genera un mensaje de error, así mismo los nombres fraudulentos son agregados a la *blacklist* por los administradores del sistema (Kaur

et al., 2012). Como las etiquetas ingresadas son libres, no escapan de poseer ruido, ser ambiguas o duplicadas. Para ello se emplea un conjunto de técnicas manuales y automáticas que resuelven el problema, además el usuario está obligado a elegir al menos una de las etiquetas predefinidas, correspondientes a la disciplina relacionada con el autor consultado, que ofrece el sistema. Las etiquetas predefinidas son las categorias del *JCR*, estas 242 disciplinas están compuesta por *Science Citation Index Expanded*, *Social Sciences Citation Index*, y *Arts and Humanities Citation Index* de *Web of Science*.

Son variadas las métricas para el recuento de citas que ofrece Scholarometer, entre ellas se encuentran h-index, g-index, h_m-index y *universal* h-index o h_f-index(Radicchi et al., 2008), el cual permite comparar cuantitativamente el impacto de los autores en diferentes disciplinas, con diferentes patrones de citas.

Para resolver el problema de los nombres de autores ambiguos existen variados esquemas como se verá en un apartado posterior. *Scholarometer* utiliza un conjunto de técnicas basadas en el análisis de las distintas variantes de los nombres de autor y en la utilización de los metadatos como los co-autores, títulos de las publicaciones, lugar de publicación y las disciplinas vinculadas entre otros, que en conjunto logran una precisión cercana al 80 %(Sun et al., 2013).

2.4.3. Microsoft Academic Search

Microsoft Academic Search (MAS) es un servicio gratuito desarrollado por *Microsoft Research* para ayudar a los investigadores, científicos, estudiantes y profesionales de forma rápida y fácil, encontrar contenido académico, investigadores, instituciones y actividades. *Microsoft Academic Search* indexa no solo millones de publicaciones académicas(al rededor de 46 millones de publicaciones y cerca de 20 millones de autores entre todas las disciplinas[17] hasta el año 2013), sino que también muestra las relaciones clave entre dos o más asignaturas, contenidos y autores, poniendo de relieve las relaciones críticas que ayudan a definir la investigación científica.

Este producto, heredero de *Windows Live Academic* y de *Live Search Academic*, surgió en su actual denominación en 2009, solamente para el

[17] http://academic.research.microsoft.com/About/help.htm

campo de la Informática, pero cubre desde septiembre de 2011 todos los ámbitos del conocimiento. Además el producto de Microsoft no se ciñe a la escala personal sino que permite seguir la pista a la producción científica de una institución e incluso efectuar comparaciones entre ellas, tomando los parámetros habituales de producción y citas como términos de la ecuación. Otra de sus fortalezas es la posibilidad de explorar la red de colaboraciones de un investigador, así como las relaciones a través de las citas. Además, se pueden encontrar los perfiles de cualquier investigador, no solo de los registrados, ya que esto no es imprescindible(Cabezas-Clavijo and Torres-Salinas, 2012).

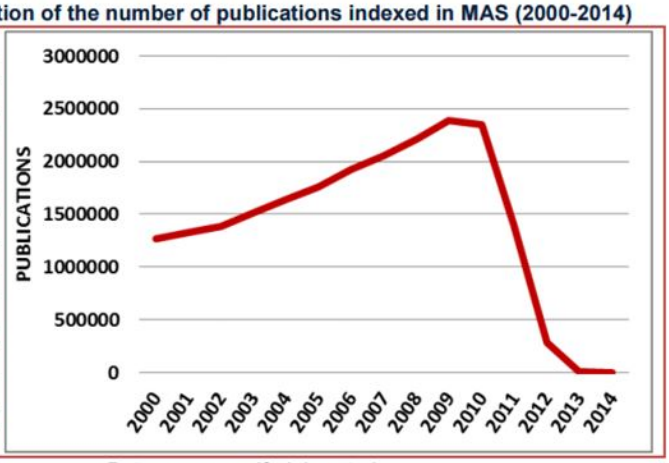
La cobertura de *MAS* frente a *GS* es mucho menor (Sun et al., 2013), si bien es un producto nuevo, la cobertura de MAS se expande rápidamente, o eso era lo que se creía en un principio. Si bien no existe abundante bibliografía que trate o que haya estudiado a MAS, un trabajo de (Orduña-Malea et al., 2014) al comparar MAS con GS en relación a cobertura y rankings de indicadores, grado de uso y visibilidad de estas bases de datos, los pobres resultados obtenidos por MAS, llevaron a un descubrimiento inesperado e inadvertido, MAS está desactualizado desde 2013. Los datos indican una caída abrupta en el número de documentos indexados, de 2.346.228 en 2010 a 8.147 en 2013.

Otro resultado obtenido de este estudio, indica que si bien MAS está desactualizado, existe un número mayor de perfiles de autores en MAS comparado con GSC (la comparación fue realizada en Agosto del 2012), el origen de este desfasaje es que los perfiles en MAS son creados automáticamente mientras que los perfiles en GSC son creados personalmente. En este sentido, existe un alto número de perfiles duplicados y desactualizados (MAS ofrece un esquema de desambiguación de nombre mediante agrupación de las publicaciones asociados a un nombre de autor y agrupación de perfiles derivados de las combinaciones en los nombres del autor). Un autor, al crear un perfile en GSC puede incluir información de afiliación, palabras claves y los documentos que desee, esto puede ocasionar la manipulación intencionada de los indicadores, como lo indica (Delgado López-Cózar et al., 2014). En la Figura 2.7 (figura tomada del trabajo de (Orduña-Malea et al., 2014)) se puede observar la evolución del número de publicaciones indexadas por MAS desde 2000 a 2014.

Otros estudios comparativos entre GS y MAS fueron los llevados a cabo por (Haley, 2014; Ortega and Aguillo, 2014), es interesante destacar los resultados experimentales realizados por (Khabsa and Giles,

Table 7. Evolution of the number of publications indexed in MAS (2000-2014)

YEAR	RECORDS
2000	1,266,245
2001	1,325,791
2002	1,385,532
2003	1,512,775
2004	1,645,212
2005	1,758,193
2006	1,922,476
2007	2,057,108
2008	2,215,580
2009	2,393,655
2010	2,346,228
2011	1,393,964
2012	290,506
2013	8147
2014	802



Data source: self-elaborated

Figura 2.7: Evolución del número de publicaciones indexadas por MAS (2000-2014)

2014) para medir el tamaño de las bases de datos bibliográficas a través de muestras aleatorias de publicaciones, luego estos documentos eran consultados por cada motor, para un artículo p se obtenía el conjunto S de artículos que citan a p, este mismo procedimiento se repetía en cada motor, luego la intersección de ambos conjuntos daba como resultado la cobertura de cada uno y mediante estimaciones se obtuvo el tamaño relativo de cada base de datos como se puede observar en la Figura 2.8 (figura tomada del trabajo de (Khabsa and Giles, 2014))

Por mucho tiempo, MAS fue un enigma, la documentación oficial de la herramienta era escasa, existía un número reducido de artículos y trabajos sobre esta herramienta, donde el principal objetivo era la comparación con GS, y en todos los casos GS mostraba su superioridad en la mayoría de los aspectos evaluados. Sin embargo, aún no se sabe desde cuándo, *Microsoft* lanzó la versión 2 de su buscador académico, ahora llamado *Microsoft Academic (MA)* [18], indicando que el proyecto *Microsoft Academic Search* (MAS) original fue finalizado en 2012 y desde entonces no se ha agregado contenido nuevo a dicha herramienta, esto responde a las cuestiones analizadas en los estudios comparativos mencionados anteriormente.

[18] http://academic.microsoft.com/

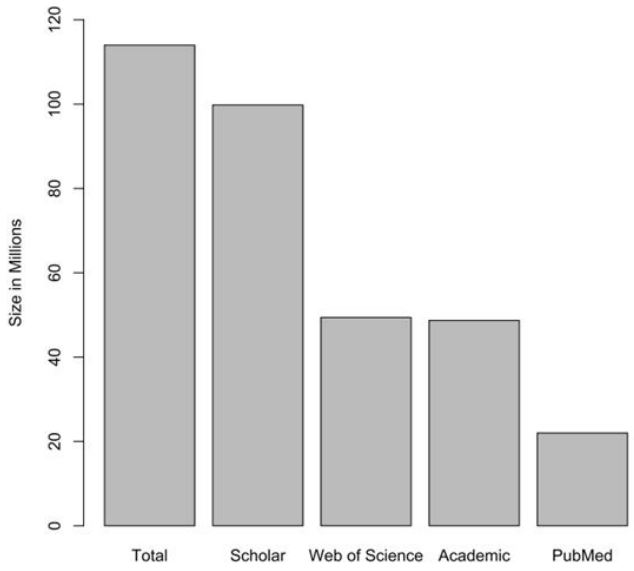


Figure 2. Relative number of documents by scholarly search engines and databases. Total and Google Scholar are estimates.
doi:10.1371/journal.pone.0093949.g002

Figura 2.8: Número relativo de documentos por motor de búsqueda académico y base de datos

Es que *Microsoft* tiene a veces esa particularidad de no informar cuándo un servicio deja de funcionar, y si no fuese por estudios experimentales como el de (Orduña-Malea et al., 2014) no se habría demostrado o verificado esa realidad un tanto oculta, de que MAS había dejado de indexar documentos. La documentación oficial de MA (Research, 2016) no indica cuando fue lanzado el nuevo servicio, el cual posee más de 80 millones de publicaciones [19] según se indica.

Este servicio es un tanto diferente del anterior, una nueva característica son las sugerencias de consultas semánticas que identifica autores, asuntos, revistas, conferencias, etc. a medida que se escribe la consulta y ofrece formas de refinar la búsqueda sobre la base de datos. El recuento de citas mostrado por publicación refleja esta estimación basada en un modelo estadístico que aprovecha las ventajas tanto de las estadísticas locales de las publicaciones individuales y las estadísticas globales de todo el grafo académico para determinar las estimaciones del número

[19] http://academic.research.microsoft.com/

de citas, en el estudio llevado a cabo por (Khabsa and Giles, 2014), se proporciona un buen ejemplo de dicha estimación estadística.

Las visualizaciones que poseía la versión anterior de este buscador (MAS) ya no se encuentran disponibles en MA ya que estaban basadas en *SilverLight*[20] el cual no es soportado por todos los navegadores, no obstante provee una API (*Academic Knowledge API*) [21] para que cada desarrollador sea capaz de obtener los datos de allí y realizar la visualización que desee. Pero esta API posee algunas restricciones, como el pago de una suscripción para el plan *Standard* como lo indica la Figura 2.9.

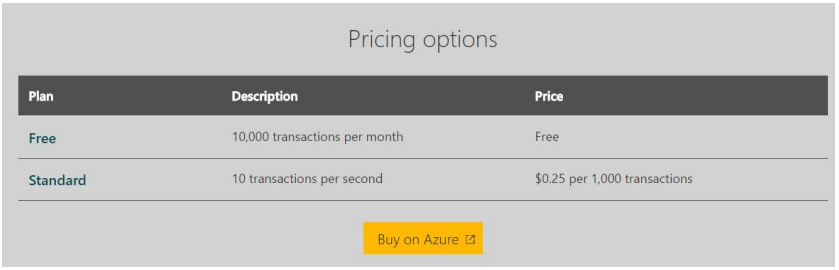


Figura 2.9: Precios para utilizar la API de Microsoft Academic

Microsft Academic viene pisando firme, y a día de hoy se presenta como una excelente alternativa a ser tomada en serio al momento de realizar análisis bibliométricos o simplemente buscar material académico. La cobertura de MA sólo es superada por GS su actual rival tanto en el número de publicaciones como en la cantidad de citas recibidas, pero algunos estudios (Harzing, 2016a; Harzing and Alakangas, 2017), limitados a ciertas áreas del conocimiento, han demostrado la superioridad, en cuanto a cobertura, de MA frente a las bases tradicionales como *Scopus* y *WoS* (ver Figura 2.10).

La popularidad del nuevo servicio de *Microsoft* viene dado no solo por la amplia cobertura demostrada, sino también por la existencia de la API que permite realizar consultas y recuperaciones de datos de forma automática con simples parámetros (en el **Capítulo 3** se explicará la inclusión de la API de MA en este proyecto).

[20] https://www.microsoft.com/silverlight/

[21] https://www.microsoft.com/cognitive-services/en-us/academic-knowledge-api

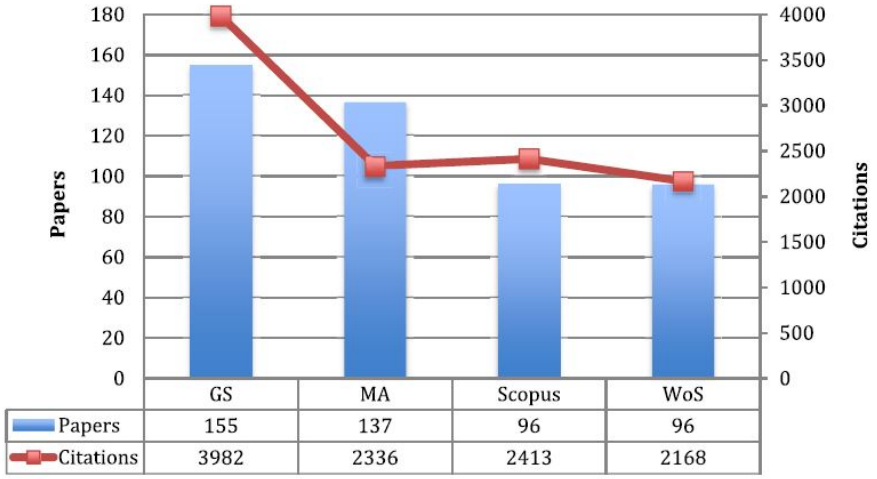


Figura 2.10: Imagen tomada de (Harzing and Alakangas, 2017) que muestra el número promedio de artículos y citas para 145 académicos en *Google Scholar*, *Microsoft Academic*, *Scopus* y *Web of Science*

2.4.4. BASE

Bielefeld Academic Search Engine (BASE) [22] es uno de los motores de búsqueda más voluminosos, multilenguaje y multidisciplinar del mundo, especialmente para recursos web académicos. A Octubre de 2016, *BASE* proporciona información de más de 100 millones de documentos provenientes de más de 4.800 fuentes. Es posible acceder a texto completo a aproximadamente el 60 % de los documentos indexados de forma gratuita (Open Access). Esta potente herramienta es operada por la Biblioteca de la Universidad de Bielefeld, Alemania (BASE, 2016a). La misma, recolecta, normaliza e indexa desde 2004 (BASE, 2016b), los metadatos de todo tipo de recursos académicamente relevantes - revistas, repositorios institucionales, colecciones digitales, etc. - que proveen una interfaz *OAI* y utilizan *OAI-PMH* para proveer sus contenidos. Asimismo BASE es un proveedor registrado de servicio OAI. Los administradores de bases de datos pueden integrar el índice *BASE* en su infraestructura local (por ejemplo, meta buscadores, catálogos de bibliotecas) (Pieper and Summann, 2015).

[22] https://www.base-search.net

En comparación con los motores de búsqueda comerciales, BASE se distingue por las siguientes características (BASE, 2016a):

- Recursos académicos seleccionados.
- Los documentos deben cumplir con requerimientos específicos de calidad y relevancia académica.
- Las búsqueda son entregadas con transparencia para el usuario, generando un inventario de recursos disponibles.
- Divulga los recursos web de la "Deep Web", que son ignorados por los motores de búsqueda comerciales o que se pierden en la gran cantidad de visitas.
- Corrección, normalización y enriquecimiento de metadatos por medio de métodos automatizados.
- La visualización de los resultados de búsqueda incluye datos bibliográficos precisos.
- Visualización del acceso y términos de reutilización de un documento.
- Varias opciones para ordenar la lista de resultados y opciones de filtrado que incluyen: (Por autor, asunto, DDC, año de publicación, proveedor de contenido, idioma, tipo de documento, acceso y condiciones de reutilización).
- Navegando por DDC (Dewey Decimal Classification), tipo de documento, acceso y términos de reutilización/licencia.

2.4.5. Google Scholar

Para hablar de *Google Scholar* sería necesario escribir un libro completo, y tal es el caso de la reciente publicación del libro "**La revolución Google Scholar**. Destapando la caja de Pandora académica" realizada por el grupo EC3, fruto de la intensa labor investigadora desplegada durante años, que como bien indica la profesora *Anne-Wil Harzing* (creadora entre tantas cosas de la genial herramienta *Publish or Perish*), en uno de los prólogos de este libro, que remarca el notable trabajo que viene llevando a cabo desde 2008, el equipo liderado por Emilio Delgado

López-Cózar, para explorar el funcionamiento interno de *Google Scholar*, enfocando sus esfuerzos en proporcionar una buena imagen sobre este motor académico (Orduña-Malea et al., 2016a).

Y si bien, otro científico prolífico en el ámbito de la bibliometría que no necesita presentación, el profesor Peter Jacsó, encargado de prologar también esta obra, que en los comienzos de *Google Scholar* se encargó de criticar y remarcar con fundamentos las falencias y errores de este motor que luego con el tiempo las fue subsanando y mejorando hasta llegar a lo que es hoy en día, subraya con el objetivismo característico de este científico, el buen y el mal uso de este motor como alternativa a las bases de datos bibliográficas tradicionales, y el cuidadoso manejo que se debe realizar sobre el mismo a la hora de obtener resultados y evaluaciones. En esta misma obra se detalla progresivamente el inicio y concepción de este motor, con lo cual se recomienda revisarla a fondo.

Desde su lanzamiento allá por 2004 hasta la actualidad, *Google Scholar* ha mejorado notablemente, la cobertura espacial se incrementó exponencialmente, tanto que nadie sabe a ciencia cierta cuál es el tamaño de esta gran base de datos, el número de registros maestros, el número de citas totales, la lista de revistas cubiertas y la cobertura temporal cubierta, salvo algunos trabajos experimentales como el llevado a cabo por (Khabsa and Giles, 2014) que han realizado algunas estimaciones del tamaño de esta base de datos. Del mismo modo que nadie duda de las bondades de esta magnífica herramienta ni de su amplia cobertura, *Google Scholar* ha sido ampliamente analizado y cuestionado tanto por la Academia como por el mundo de los profesionales de la información, evaluando, analizando y comparando tanto sus fortalezas como sus debilidades. Si bien el debate continua, el uso adecuado y responsable de esta herramienta es una opción más que válida.

Otro factor que también ha ido en aumento es la popularidad de *Google Scholar*, y el amplio rango etario y niveles académicos que lo utilizan. Ya no se habla solamente de estudiantes universitarios utilizando este motor para hacer consultas de material académico, sino también de investigadores, científicos, estudiantes de doctorado, post-doctorales, profesores e incluso bibliotecarios. La tendencia marcada en la mayoría de los casos es iniciar la búsqueda en *Google Scholar* (alentada esta conducta principalmente por ser gratuito, fácil de usar, intuitivo, de gran alcance, y multidisciplinario) para consultar cierto material bibliográfico y solo en caso de ser necesario y posible, acceder a las bases de datos tra-

dicionales para obtener ciertos datos no aportados por *Google Scholar*, como los metadatos y datos estadísticos más precisos.

Como dato al margen, este trabajo posee un doble enfoque hacia *Google Scholar*, por un lado como herramienta de búsqueda de información, el 100 % de la bibliografía fue consultada y extraída de este motor, salvo algunos artículos particulares donde no existía el texto completo (no eran publicaciones *Open Access*) y se utilizó la suscripción de la Universidad de Salamanca a revistas y bases de datos para obtener el texto completo. Y por otro lado como ya resulta obvio, como objeto de investigación y fuente de datos para los resultados a analizar posteriormente.

Google Scholar como motor de búsqueda indexa todo lo que considera "contenido académico", para ello cuenta con una seria de robots o *crawlers* encargados de detectar este tipo de recursos disponibles en la Web pública, si bien se entiende que *Google Scholar* posee su propio sistema de robots encargados de recolectar este tipo de material, los robots del buscador *Google* de propósito general son diferentes, pero existe cierta sinergia entre estos. Como lo indica (Orduña-Malea et al., 2016a), *Google Scholar* solamente indiza "documentos académicos" alojados en la "Web académica" (Figura 2.11 tomada de (Orduña-Malea et al., 2016a) Capítulo 4 - Capturando la Web académica). La cuestión ahora radica en conocer qué entiende exactamente *Google Scholar* por documento académico y por Web académica, y cómo realiza a grandes rasgos este proceso.

El sistema se basa fundamentalmente en la estructura del documento y en la posterior identificación de cada una de sus partes constituyentes. Aunque existen diversos métodos más o menos implantados para estructurar un documento científico (como por ejemplo el formato IMRYD para el caso de los trabajos de investigación, basado en el siguiente orden de elementos: Introducción, Metodología, Resultados y Discusión), no existe ninguna norma obligatoria para determinar el tipo, nomenclatura y orden exacto de los elementos constitutivos de un documento científico (más allá del sentido común), siendo un aspecto delimitado generalmente por las normas específicas de una publicación.

Pese a ello, existen ciertos elementos que deben aparecer en cualquier texto que se considere académico, con independencia de la estructura organizativa utilizada, como son un título, unos autores o una bibliografía, entre otros. Siguiendo este principio, *Google Scholar* considera los si-

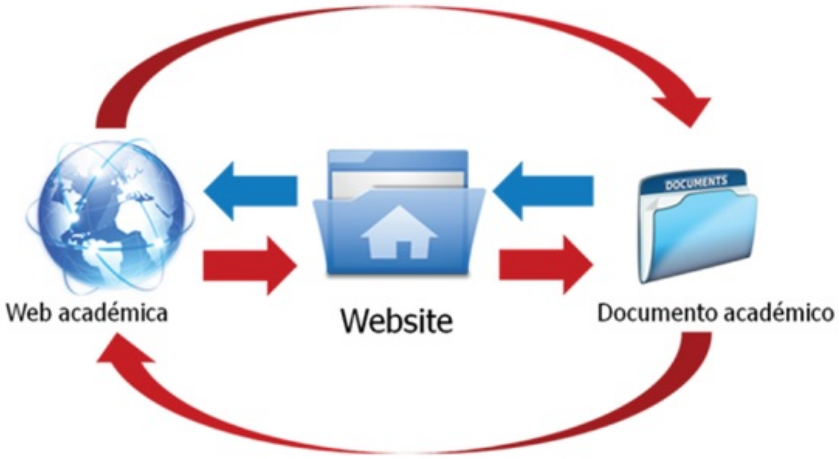


Figura 2.11: Indización de contenidos en Google Scholar

guientes aspectos

- Que el documento esté encabezado por un título (a ser posible en una fuente mayor que el resto del documento).
- Que posea unos autores (deben mostrarse justo debajo del título en una línea separada).
- Que tenga resumen.
- Que contenga una sección independiente titulada como Bibliografía, referencias, o similar.

Una vez determinado qué es un documento académico para *Google Scholar*, el segundo requisito para la incorporación de un documento a la base de datos es que este recurso esté alojado en la llamada "Web académica". En este punto, *Google* simplemente confía en los espacios públicos web proporcionados por diferentes entidades y organizaciones con cierto marchamo científico o académico. Entre éstas destacan las siguientes entidades:

- Universidades e instituciones de educación superior.

- Organismos públicos de investigación.
- Repositorios temáticos.
- Editoriales comerciales.
- Servicios de distribución o almacenaje.
- Bases de datos bibliográficas.
- Otros motores de búsqueda académicos.

En cualquier caso, nada impide que otro tipo de institución (pública o privada), pueda estar indizada correctamente en *Google Scholar*. Adicionalmente, se debe indicar la existencia de acuerdos comerciales no públicos entre *Google Scholar* y ciertas editoriales privadas. En estos casos, el buscador incorpora a su base de datos los contenidos correspondientes.

2.4.6. Iniciativas altmetrics

Han pasado 351 años (1665) desde la primer publicación de las revistas *Journal des Sçavans* y *Philosophical Transactions*, marcando el nacimiento de la revisión por pares de artículos de revista. Esta forma de comunicación académica, no sólo se ha mantenido como el modelo dominante para la difusión de nuevos conocimientos (sobre todo para la ciencia y la medicina), sino que también ha aumentado considerablemente en volumen(Haustein et al., 2015b).

Durante los últimos 50 años, el análisis de citas y, en general, los métodos bibliométricos, se han desarrollado a partir de herramientas de recuperación de información para investigar las métricas de evaluación, bajo la suposición de lograr que la financiación científica sea más eficiente y eficaz (Moed, 2006). Las diferentes áreas del conocimiento están pobladas por comunidades de científicos y profesionales, cada grupo utilizando sus propias herramientas, metodologías y técnicas. Estos son los grupos sociales que comparten, con más o menos consenso, prácticas profesionales, las formas de organización del trabajo, condiciones de vida, expectativas sociales, principios, valores y creencias (Martín-Martín et al., 2016c).

La sola existencia de éstas métricas para evaluar la investigación científica ha creado efectos adversos como la locura o fiebre por publicar a como dé lugar, o la existencia del *web spam* como se indicó anteriormente y los problemas que esto conlleva, especialmente con los motores de búsqueda de acceso libre. Con el nacimiento de la web social, la comunicación científica se está volviendo cada vez más abierta, transparente y diversa. En números crecientes, los académicos están trasladando sus trabajos diarios a la web y logrando así una mayor visibilidad y transferencia de conocimiento. Las métricas alternativas permiten a los investigadores conocer en menor tiempo la repercusión de sus trabajos (Torres-Salinas and Milanés-Guisado, 2014), al poder publicar en los medios digitales sin tantas restricciones y procesos metodológicos como una revista, nuevas medidas e indicadores pueden obtenerse en tiempo real y de varias fuentes de datos al mismo tiempo.

Además de las bases de datos tradicionales para evaluar la productividad científica y además de los distintos indicadores mencionados, existen otro tipo de iniciativas dirigidas a medir el impacto de los materiales publicados online, de forma más amplia que con el recuento de citas. Estas iniciativas son conocidas como *altmetrics* o *métricas alternativas*(Priem et al., 2010). El término *altmetrics* fue introducido por primera vez en un *Tweet* en 2010 (ver Figura 2.12) y más tarde (Priem, 2014) amplió la definición para incluir medidas de impacto académicas disponibles en cualquier plataforma en línea.

El conjunto de métricas comúnmente conocida como *altmetrics*, se basan generalmente en la medición de la actividad en línea relacionada con los académicos o los contenidos académicos derivados de las redes sociales y plataformas web 2.0. Sin embargo, la definición de lo que constituye un indicador *altmetrics* está en constante cambio, ya que está determinada en gran medida por las posibilidades técnicas y, más específicamente, la disponibilidad de Interfaces de Programación de Aplicaciones (API) (Haustein et al., 2015b). Si bien no existe un consenso o no todos los autores están de acuerdo con el termino "altmetrics", el denominador común de varios *altmetrics* es que excluyen y se oponen a los indicadores bibliométricos "tradicionales".

La idea que subyace es que, por ejemplo, las menciones en *blogs*, el número de *retweets* o el de personas que guardan un artículo en su gestor de referencias puede ser una medida válida del uso de las publicaciones científicas. Sin embargo, la medición de la visibilidad de la ciencia en In-

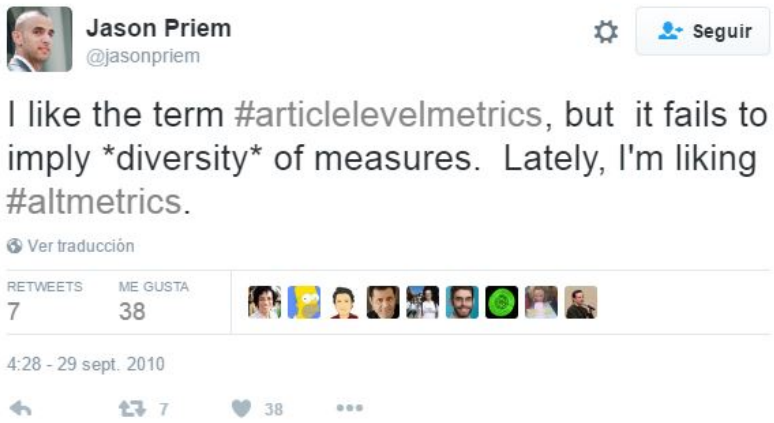


Figura 2.12: Tweet de Jason Priem haciendo mención por primera vez al término altmetrics

ternet no es un nuevo fenómeno. El origen de las *altmetrics* se remonta a los años 90 con la webmetría, el estudio cuantitativo de las características de la web (Thelwall et al., 2005), que nace de la aplicación de las técnicas bibliométricas a los sitios online, y engloba diversas disciplinas, entre ellas, la comunicación (Torres-Salinas et al., 2013).

(Haustein et al., 2015b,a) provee un panorama muy claro y profundo desde el origen de *altmetrics* hasta la actualidad, abordando temas como la cobertura general, la cobertura por disciplina, las posibilidades que estas herramientas proveen, la fiabilidad y validez de las métricas de medios sociales, el interés público en la ciencia, el sesgo geográfico, el tipo de material citado en los medios sociales (artículos que reciben al menos un *tweet*, compartidos en *Facebook*, mencionado en blogs, encontrados en *Google+* y discutidos en los medios tradicionales), entre otros temas.

En la Figura 2.13 (imagen tomada del estudio llevado a cabo por (Torres-Salinas et al., 2013)), se puede observar una lista bien clasificada de las principales medidas propuestas por las *altmetrics*.

También conocidas como Redes Sociales Digitales Científicas (RSDC). Se pueden definir como plataformas de comunicación en red, que posibilitan a los investigadores crear y dar a conocer un perfil académico y profesional. Este perfil es la puerta de acceso para difundir investiga-

Tabla 1. Principales medidas propuestas por las altmetrics clasificadas según el tipo de plataforma, indicador y red social o plataforma

Tipo de plataforma	indicadores	Red social o plataforma	Ejemplos de indicadores
BIBLIOTECAS Y GESTORES DE REFERENCIAS DIGITALES	Social bookmarking y biblioteca digitales	Generales: - Delicious Académicas: - Citeulike - Connotea - Mendeley	Nº de veces que ha sido favorito Nº de lectores Nº de grupos a los que se ha añadido
REDES Y MEDIOS SOCIALES	Menciones en redes sociales	Generales: - Facebook - Google+ - Twitter Académica: - Academia.edu - Research Gate	Número de me gusta Numero de clicks Número de comentarios Número de veces compartido Número de tuits que mencionan Número de Retwits Retwits de usuarios líderes
	Menciones en blogs	Generales: - Blogger - Wordpress Académicos: - Nature Blogs - Postgenomic blog - Research Blogging	Número de citas en blogs Comentarios a la entrada del blogs Sistemas de rating de la entrada
	Menciones en enciclopedias	- Wikipedia - Scholarpedia	Citas en entrada de las enciclopedias
	Menciones sistemas de promoción de noticias	Generales: - Reddit - Meneáme Académicas: - Faculty of 1000	Número de veces en la portada Número de Clicks (meneos) Número de comentarios a la noticias Puntuación de los expertos

Figura 2.13: Clasificación de las principales medidas propuestas por las altmetrics

ciones y consultar *online* y/o descargar referencias y otras producciones científicas. Según (González-Díaz et al., 2015), las dos principales RSDC son *ResearchGate* y *Academia.edu*.

Estas iniciativas siguen el camino trazado por las estadísticas a nivel de artículo implementadas por la editorial *Public Library of Science (PLoS)*, que lleva tiempo ofreciendo, además de las citas que un artículo atrae, datos sobre el número de descargas de un trabajo, el número de comentarios que genera o el número de *blogs* que lo enlazan(Cabezas-Clavijo and Torres-Salinas, 2010).

Ejemplo del uso de éstas herramientas es el análisis conducido por (Martín-Martín et al., 2016c) en el que se analiza en detalle la presencia de un conjunto de investigadores, una muestra de 814 autores con perfiles públicos en GSC (de estos autores, el *Core* formado por 398 autores cuya producción científica cae sustancialmente dentro del campo *Bibliometrics*, y el *Related* 416 autores que han publicado estudios bibliométricos de forma esporádica, o cuyo campo de especialización está estrechamente relacionado con *Scientometrics*, y por lo tanto no pueden ser considerados estrictamente bibliométricos). La idea es comparar la presencia de este conjunto de investigadores en las plataformas sociales o perfiles utilizados por los académicos (*ResearcherID*, *ResearchGate*, *Mendeley* y *Twitter*). De mayor a menor presencia se ubicaron: GSC (por obvias razones ya que se partió de la existencia del perfil público en esta plataforma), *ResearchGate*, *Mendeley*, *ResearcheID*, y *Twitter* (en la Figura 2.14 se observan estos resultados). Solo el 11 % de la muestra (93 autores) poseen perfiles creados en todas las plataformas mencionadas. La idea de este análisis era revelar las relaciones entre las métricas y la plataforma analizada, los resultados obtenidos se pueden separar en dos clases de impacto en la web, el primero en cuanto a las métricas relacionadas con el impacto académico: métricas de uso (vistas y descargas) y métricas de citas. Y el segundo relacionado con las métricas de conectividad y popularidad (seguidores).

Table 7. Degree of use of social platforms by type of author

WEB PLATFORMS	AUTHORS CORE	%	RELATED	%	TOTAL	%
* Google Scholar Citations	398	100	416	100	**814**	100
ResearcherGate	260	65.33	283	68.03	**543**	66.71
Mendeley	171	42.96	165	39.66	**336**	41.28
** Homepage	158	39.69	177	42.54	**335**	41.15
ResearcherID	182	45.73	146	35.10	**328**	40.29
Twitter	132	33.17	108	25.96	**240**	29.48

* All authors in the sample have a profile in GSC. ** *ResearchGate* and *Academia.edu* URLs were discarded.

Figura 2.14: Utilización de cada una de las plataformas sociales

Estas nuevas formas reflejan y transmiten el impacto académico: ahora las publicaciones pueden ser accedidas, rastreadas, y sus accesos contabilizados. Este grupo diverso de actividades forman una traza compuesta del impacto, mucho más rica que cualquiera de sus predecesoras. *Altmetrics* amplía no solo la visión de lo que parece representar el impacto,

sino también de lo que hace al impacto. Esto es importante porque las expresiones del conocimiento son cada vez más diversas(Priem et al., 2010).

A pesar de que aun las publicaciones impresas siguen teniendo una fuerte influencia en la comunidad académica, los medios sociales como *blogs*, repositorios, redes sociales y gestores de referencias *online* están empezando a ser considerados con el objetivo de obtener una imagen más completa acerca del impacto de las publicaciones. Tanto las citas (análisis basados en el recuento de citas) como las métricas de medios sociales aumentan con el grado de colaboración y la longitud de la lista de referencias. Existen materiales muy pocos citados pero que son populares en medios sociales, y en algunas disciplinas sucede lo contrario. Estos resultados sugieren que los factores que impulsan los medios sociales y las citas son diferentes. En este sentido, las métricas de medios sociales no pueden ser vistos como alternativas a las citas(Haustein et al., 2015a), es decir, los indicadores *altmétricos* complementan o mejoran los sistemas de evaluación científica tradicionales (Robinson-García et al., 2014).

2.4.7. Herramientas altmetrics

A continuación se ofrece una breve introducción de las principales herramientas altmetrics utilizadas hoy en día por la comunidad investigadora, existen algunas creadas con un propósito específicamente académico y hay otras que fueron adaptándose o la misma comunidad las fue incorporando en su labor diaria par hacer visible los resultados de la investigación.

2.4.7.1. ResearcherID

ResearcherID[23] lanzada por *Thomson Reuters* en 2008, es una plataforma para administrar y compartir información profesional, resuelve los problemas de autoría al mismo tiempo que añade, al perfil personal, métricas de citas dinámicas (tomando como fuente de datos la Web of Science) y redes de colaboración. *ResearcherID* ofrece a la comunidad mundial de investigación un índice de valor incalculable de la información del autor. Al asignar un identificador único a cada autor participante,

[23] http://www.researcherid.com/

ResearcherID estandariza y clarifica los nombres de autor y las citas, logrando que la información de búsqueda sea más sencilla y accesible. Debido a que *ResearcherID* está totalmente integrado con la *WoS*, se puede utilizar el perfil para asegurarse de que un autor está siendo debidamente acreditado por su trabajo en la *WoS*, administrar la lista de publicaciones, rastrear las citas a través del tiempo y obtener indicadores (h-index), identificar posibles colaboradores y utilizar *ResearcherID* para encontrar el cuerpo de la obra de un autor. *ResearcherID* también permite asociar el ORCID del autor a la cuenta del autor (Reuters, 2016a).

2.4.7.2. ResearchGate

ResearchGate[24] fundado en 2008, es una red social y una herramienta colaborativa dirigida a la comunidad investigadora, donde se puede crear un perfil propio, listar las publicaciones e interactuar con otros usuarios (Thelwall and Kousha, 2015). El año pasado alcanzó los 8 millones de investigadores y científicos (ResearchGate, 2016). *ResearchGate* ha permitido que los investigadores difundan sus ideas y compartan sus publicaciones de forma gratuita, de modo tal de facilitar la colaboración entre los investigadores de todo el mundo. A través de *ResearchGate*, los usuarios pueden utilizar la plataforma para mantener sus propias publicaciones, generar y responder a preguntas relacionadas con la investigación, y seguir a otros investigadores para recibir las actualizaciones de sus publicaciones. *ResearchGate* intenta combinar ambas bibliometrics y altmetrics para crear una medida más amplia de rendimiento tanto para instituciones como investigadores.

Las métricas tradicionales al igual que un indicador de rendimiento, tienen como objetivo medir la cantidad y calidad de publicaciones como libros y artículos. En *ResearchGate*, cuatro indicadores bibliométrico (puntos de impacto, el número de publicaciones, número de descargas y visitas al perfil) se utilizan para medir el impacto de las instituciones académicas e investigadores. La puntuación *ResearchGate* (ResearchGate score), un nuevo indicador de referencia, calculado por un algoritmo aun no revelado, se utiliza para integrar ambas *bibliometrics* y *altmetrics* mediante la medición de las publicaciones del investigador, las preguntas formuladas y contestadas, y el número de seguidores.

[24] https://www.researchgate.net/

Por lo tanto, *ResearchGate* afirma que su puntuación mide la reputación científica sobre la base de las contribuciones individuales de un investigador, las interacciones, y la reputación (Yu et al., 2016). El sistema se basa fundamentalmente en la capacidad para depositar y almacenar cualquier documento académico por parte de los autores (desde un artículo publicado en una revista de impacto hasta patentes, comunicaciones a congresos, materiales de un curso, una presentación o *datasets*) y en la inmediata obtención de estadísticas de uso personalizadas (quién visita, descarga o cita un documento o a sus autores). A fecha de febrero de 2016, la plataforma informa de la disponibilidad de más de 81 millones de publicaciones (ver Figura 2.15, de las que aproximadamente el 23,5 % se encuentran a texto completo), con una representación de 193 países y con miembros verdaderamente insignes, entre los que se encuentran 52 investigadores galardonados con el premio Nobel (Orduña-Malea et al., 2016b).

Tabla 1. Tamaño de las principales bases de datos bibliográficas (marzo de 2016)

Base de datos	Número de documentos
*Google Scholar***	170.000.000*
Web of Science (todas las bases de datos)	167.127.889
ResearchGate	**81.000.000***
Microsoft Academic Search	80.000.000*
Web of Science Core Collection	61.856.513
Scopus	61.583.942
Mendeley	32.000.000*
Academia.edu	10.767.769*

* datos aproximados
** datos a fecha de junio de 2014

Figura 2.15: Estimaciones del tamaño de las principales bases de datos bibliográficas

2.4.7.3. Mendeley

Mendeley[25] es un gestor de referencias bibliográficas gratuito y una red social académica que permite organizar la investigación, colaborar con otros en línea, y descubrir las últimas investigaciones[26]. Lanzado en agosto de 2008, es uno de los más populares *ASNSs (Academic Social Networking services)* pero se distingue por permitir a los usuarios formar grupos de interés (Jiang and Gao, 2016). Posee más de dos millones de usuarios. *Mendeley* permite a los usuarios crear su propia biblioteca digital de investigación mediante la importación de archivos PDF desde sus dispositivos locales. Luego que un artículo es añadido a la biblioteca, se extraen automáticamente los metadatos, incluyendo el título, autores, año de publicación, el nombre de la revista o acta, y así sucesivamente. Los usuarios son capaces de abrir los archivos PDF importados, comentarlos, tomar notas y resaltar el texto. Una vez que se añadió un artículo a la biblioteca de un individuo, todos los usuarios son capaces de buscar el documento en el catálogo *Mendeley*.

A diciembre de 2012, el catálogo en línea de *Mendeley* contenía más de 300 millones de trabajos de investigación. Al igual que con un software de referencia común (por ejemplo, EndNote [27]), Mendeley ofrece capacidades para gestionar y generar datos de citas. Los usuarios pueden crear bibliografías, ya sea a partir de los documentos que se agregan o desde el catálogo en línea. Hay tres formas comunes de utilizar las funciones sociales en *Mendeley*: mantener un perfil, gestionar los contactos existentes y hacer más conexiones. El "perfil" es una característica común de la mayoría de los sitios de redes sociales. *Mendeley* permite a los usuarios crear un perfil profesional con propiedades orientadas a la investigación (pueden listar las publicaciones, áreas de interés, premios y becas en su propia página de perfil). *Mendeley* también permite a los usuarios iniciar grupos para compartir intereses en común y lo que están leyendo.

Existen dos tipos de grupos soportados por el sitio: grupos privados que son accesibles solamente a los miembros y grupos públicos que son visibles públicamente y se pueden buscar en la lista de grupos de *Mendeley*. Los propietarios de los grupos públicos pueden decidir si la

[25] https://www.mendeley.com/
[26] https://www.elsevier.com/solutions/mendeley
[27] http://endnote.com/

adhesión es accesible a todos los usuarios o si tiene que ser revisado y aprobado. Por lo tanto, los grupos públicos son totalmente abiertos o con aprobación. Los usuarios pueden participar libremente grupos totalmente abiertas, pero necesitará la aprobación del propietario para unirse a grupos con aprobación. Los usuarios pueden optar por participar en un grupo, ya sea uniéndose al grupo como miembro o siguiendo (*following*) al grupo como un seguidor (*follower*). La única diferencia entre estos dos roles es que los miembros tienen derecho a contribuir con artículos. Ambos tipos de usuarios son capaces de remarcar un artículo como un "me gusta" (*like*) y hacer otros tipos de comentarios sobre el "muro" (*wall*) del grupo (Jeng et al., 2015).

2.4.7.4. Academia.edu

Academia.edu[28] es un sitio web donde los investigadores pueden publicar sus artículos, y hallar y leer artículos escritos por otros. Combina la función de archivador como un repositorio, como es el caso de *ArXiv* [29], SSRN [30], o *PubMed* [31], con características de redes sociales, tales como perfiles, canales de noticias, recomendaciones, y la capacidad de seguir a individuos y temas de preferencias. El sitio fue lanzado en 2008, y actualmente cuenta con casi 43 millones de usuarios registrados quienes han subido cerca de 15 millones de artículos (Academia.edu, 2016). La inscripción en el sitio es gratuita y los usuarios pueden descargar libremente todos los documentos publicados en el mismo.

Academia.edu tiene características únicas para descubrir artículos, por lo que es un lugar interesante para analizar las ventajas de citar artículos. Los usuarios son notificados cuando los autores que ellos están siguiendo, publican artículos en el sitio. A continuación, pueden compartir estos artículos con sus seguidores. Un usuario puede etiquetar un artículo con un tema como "High Energy Physics" y los usuarios que "siguen" ese tema, serán notificado sobre el artículo (Niyazov et al., 2016). *Academia.edu* juega un papel en la comunicación académica formal, ya que los autores pueden subir pre-impresos y otros documentos a su perfil.

Aunque la cantidad de citas que recibe *Academia.edu* es menor a un

[28] https://www.academia.edu/
[29] https://arxiv.org/
[30] https://a1www.ssrn.com/
[31] https://www.ncbi.nlm.nih.gov/pubmed/

tercio de lo que recibe *Facebook*, con exclusión de las citas a páginas generales, su contenido se cita más que las de los otros sitios académicos especializados. *Academia.edu* permite a sus miembros listar libros, charlas, ponencias e intereses (una lista de palabras clave) en su página de perfil, junto con su nombre, una información de imagen y la afiliación. Otros usuarios, consultando el perfil del miembro, podrán observar esta información, así como el número de veces que el perfil ha sido visto, al igual que el número de veces que cada documento que se indica en el perfil ha sido visto (y un recuento general del documento)(Thelwall and Kousha, 2014)

2.4.7.5. CiteULike

CiteULike[32] es un servicio gratuito para la gestión y el descubrimiento de referencias académicas. Posee más de 8 millones de artículos indexados. *CiteUlike* permite fácilmente almacenar referencias que puedes encontrar en Internet, descubrir nuevos artículos y recursos, recomendaciones de artículo automatizada, compartir referencias con sus compañeros, descubrir quién está leyendo lo que está leyendo, y almacenar y buscar sus archivos PDF (CiteUlike, 2016). El sitio web se puede utilizar para buscar referencias públicas de todos los usuarios o sólo las propias referencias. Estas referencias se pueden exportar a través de *BibTeX*[33] o *EndNote* para ser utilizado en equipos locales.

CiteULike es un sistema abierto a cualquier persona que desea conservar, gestionar y compartir referencias de documentos científicos y técnicos en internet, que se conservan como propias pero que están visibles para todos, mediante un perfil público *CiteULike* y uno privado *MyCiteULike*. CiteULike permite crear grupos de investigación en torno a un tema o un departamento con la finalidad de compartir las referencias entre sus miembros. *CiteUlike* hace dos cosas bien: la recopilación y el descubrimiento de referencias. La parte de recuperación es la que el marcador de favoritos ofrece, una forma sencilla y eficiente para guardar el enlace y los metadatos de cualquier referencia que se puede encontrar en Internet. Y todo se almacena en línea, lo cual tiene muchas ventajas.

Cada artículo, cada biblioteca de usuarios, cada vista de la base de datos está disponible como una URL. Se puede acceder a él desde cual-

[32] http://www.citeulike.org/
[33] http://www.bibtex.org/

quier ordenador. Es muy sencillo para enlazar a partir de otras aplicaciones. La parte interesante, es el descubrimiento social que está habilitado por los usuarios que mantienen sus referencias en público. Gracias a la estructura social que ya existe en la base de datos, se ofrece al usuario variados enfoques basados en la actividad de los miembros, por ejemplo: ¿Quién más está interesado en el mismo artículo que uno?, ¿Qué artículos poseen etiquetas en común?, ¿Qué artículos están siendo marcados con frecuencia como favoritos este mes? (SocietyZone, 2016)

2.4.7.6. Altmetrics.com

Altmetrics.com[34] es una *start-up* que se utiliza bajo suscripción lanzada en julio de 2011. Muestra el impacto de la investigación a sus autores y lectores en un modo muy visual por medio de los *Donut altmetrics* (dona o rosquilla como se puede observar en la Figura 2.16) y a traves de la *Altmetric Attention Score* (puntuación de atención), que es el número indicado en el centro de la *Donut* . Tanto la *Altmetric Attention Score* como la *Donut altmetrics*, están diseñados para ayudar a identificar fácilmente qué cantidad y qué tipo de atención ha recibido un resultado de investigación. Al hacer click en la dona se puede visitar la página de detalles de los resultados, para ver las menciones originales y las referencias que han contribuido a la puntuación (la atención puede ser tanto positiva como negativa).

Esta métrica permite entre tantas cosas monitorear, buscar y medir todas las conversaciones acerca de los artículos de una revista, así como los publicados por sus competidores. En cuestión de minutos, permite al autor disponer de los datos *Altmetrics* para insertarlos y mostrarlos en su plataforma o aplicación (Arévalo and Vázquez, 2016). *Altmetric.com* es un proveedor abierto y transparente, que ofrece datos de calidad sobre menciones a publicaciones científicas en medios sociales (Robinson-García et al., 2014). Recoge las menciones de artículos académicos de todas partes de la Web mediante la recopilación de menciones en los periódicos, blogs, redes sociales y otros sitios Web tales como: fuentes de Documentos de Política, Noticias, *Blogs*, *Twitter*, post-publicaciones de revisiones por pares, *Facebook*, *Sina Weibo*, *Wikipedia*, *Google+*, *LinkedIn*, *Reddit*, *Faculty1000*, *Q&A(stack overflow)*, *Youtube* y *Pinterest* (Altmetric.com, 2016). Hoy en día la base de datos de *Altmetric* contiene

[34] `http://www.altmetric.com`

menciones de más de 4 millones de fuentes de investigación (incluyendo artículos de revistas, bases de datos, imágenes, documentos, informes y más), y está en constante crecimiento (Arévalo et al., 2016).

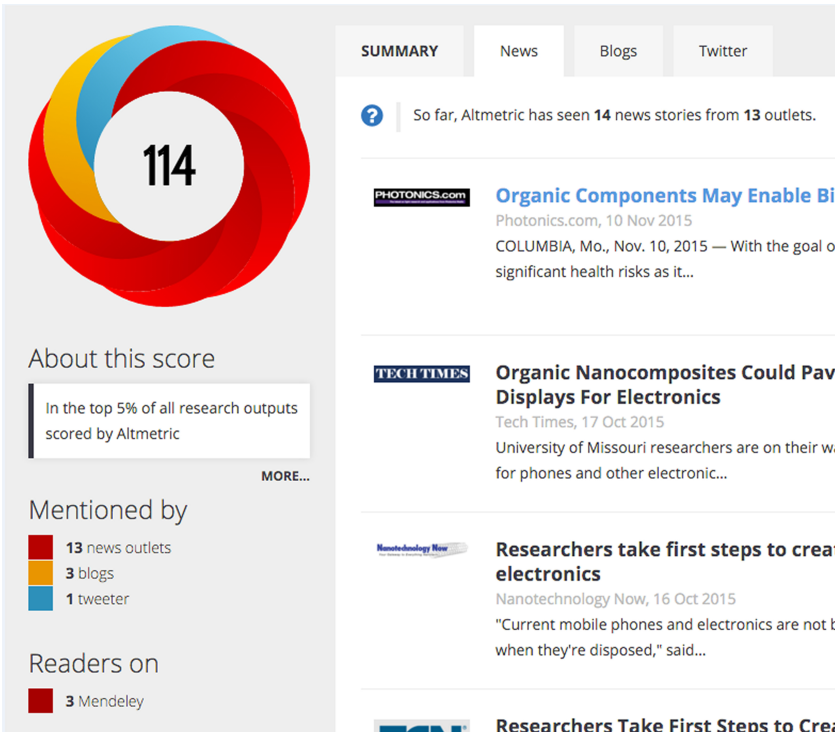


Figura 2.16: Ejemplo de Donut altmetrics

Cada fuente posee un color distinto y la cantidad de cada color en la dona cambiará dependiendo de la atención puesta en las fuentes de producción de investigación. El algoritmo *Altmetric.com* calcula una puntuación global ponderada teniendo en cuenta el volumen, la fuente y el autor en función de las menciones que recibe un documento.

En conjunto, la puntuación calculada representa una aproximación ponderada de toda la atención que se ha recogido para un resultado de investigación (no un total en crudo del número de menciones). Esta ponderación es relativa a la fuente que menciona o cita el artículo o publicación, por ejemplo, las menciones en un periódico importante como

el New York Times contribuye mucho mas que un post en Facebook [35]

La puntuación de la atención es útil cuando se miran varias salidas en conjunto para identificar rápidamente el nivel de actividad en línea que rodea a un resultado de investigación en particular (no es una medida de la calidad de la investigación, o el investigador).

Es importante tener en cuenta que los lectores *Mendeley*, número de citas *Scopus* y marcadores *CiteULike* no cuentan para la puntuación, esto se debe a que no se pueden mostrar todos los detalles de quién en realidad está haciendo la mención o referencia. La política del sitio web indica que cualquier mención que cuenta para la puntuación debe ser totalmente transparente y plenamente visible en la página de detalles *altmetrics* (Altmetric.com, 2016).

Una muestra de la aceptación y utilidad de esta poderosa herramienta es la inclusión de la dona *Altmetric.com* en *Scopus* desde junio de 2012 en la barra lateral de las páginas de resumen y del documento. Se puede observar en la parte derecha de la pantalla cuando se dispone de datos para el artículo que se está viendo. Los visitantes pueden hacer clic en la dona para revisar la mención y hacer clic en cualquier entrada para navegar hasta el sitio original. Una pestaña "demografía" también muestra un desglose de que parte del mundo está viniendo la atención para dicho artículo. Los clientes y los usuarios han encontrado con este agregado, un complemento útil de las citas tradicionales. El principal punto de interés no ha sido necesariamente las métricas en sí, sino el contenido subyacente. Al descubrir que un *blogger* de ciencia respetado ha dado una opinión positiva de un artículo es mucho más importante que saber cuántas personas han blogueado sobre él (Elsevier, 2016a).

2.4.7.7. Plum Analytics

Plum[36] (que significa ciruela) es el portal de métricas de investigación de *Ebsco*. Realiza un seguimiento de todos los productos de la investigación en cualquier forma, proporcionando una poderosa herramienta que aumenta la capacidad de la métrica tradicional. La herramienta *PlumX* reúne a las métricas a través de cinco categorías: citaciones, uso, men-

[35] https://help.altmetric.com/support/solutions/articles/6000060969-how-is-the-altmetric-score-calculated-

[36] http://plumanalytics.com/

ciones, capturas recogidas de los datos de proporcionados por los medios sociales. Los documentos incluyen: artículos, entradas del blog, capítulos de libros, libros, casos, ensayos clínicos, comunicaciones a congresos, conjuntos de datos, cifras, subvenciones, entrevistas, cartas, medios de comunicación, patentes, posters, presentaciones, código fuente, tesis / disertaciones, vídeos, páginas web. *Plum* proporciona datos objetivos sin establecer ponderaciones o ranking como *altmetric.com*. También es posible añadir *widgets* de *PlumX* a su repositorio institucional, perfiles de investigadores, sitios web del departamento, o blogs, etc (Arévalo et al., 2016).

2.4.7.8. ImpactStory

ImpactStory [37] es una organización sin fines de lucro que ayuda a los científicos a conocer, donde su investigación está siendo citada, compartida, almacenada y mucho más. Fue creado en 2011 por Jason Priem y Heather Piwowar como "Total Impact" y renombrada en 2012 como ImpactStory, está financiado por la Fundación Nacional para la Ciencia y la Fundación Alfred P. Sloan una organización no lucrativa [38].

ImpactStory es una herramienta basada en la web de código abierto que ayuda a los investigadores a explorar, dar visibilidad y compartir los diversos impactos de todos los productos de investigación, desde los tradicionales como libros y artículos de revistas, hasta los productos emergentes como blogs, bases de datos, diapositivas(slides) y software. Es una herramienta altmétrica que además de proporcionar a los investigadores datos de su impacto, está ayudando a construir un nuevo sistema de reconocimiento académico que valora y fomenta la repercusión de la investigación en la web (Arévalo and Vázquez, 2016).

El perfil se crea a partir de la introducción de ítems, con las siguientes restricciones: los documentos deben disponer de DOI y los documentos deben estar incluidos en ORCID.

[37] https://impactstory.org/
[38] https://impactstory.org/about

2.4.7.9. PLoS Article Level Metrics (ALM)

Public Library of Science (PLoS) se ha convertido en el principal repositorio de revista de acceso abierto, en parte debido a sus tradicionales altos índices de impacto. Sin embargo, PLoS ofrece una alternativa gratuita al impacto tradicional en forma de *Article Level Metrics (ALM)*[39] (métricas a nivel de artículo), que siguen la influencia de los artículos *PLoS* individuales, desde las veces que son descargados hasta las menciones en medios sociales y blogs. ALM se implementó en 2009 y realiza el seguimiento del alcance, el uso y la reutilización de los resultados de la investigación (desde los artículos y figuras a los conjuntos de datos y el código) para ayudar a guiar la comprensión de la influencia de una obra. PLoS también rastrea las métricas internas del artículo, incluyendo los comentarios, notas y calificaciones. Si bien esto representa un recurso valioso del impacto, sólo los artículos PLoS se benefician de sus métricas (Roemer and Borchardt, 2012).

Con ALM, se puede ver una colección de indicadores de impacto en tiempo real. Lo cual permite al investigador mantenerse al día del alcance y la influencia de su investigación y, a continuación, compartir esta información con sus colaboradores, el departamento académico, y sus financiadores. ALM tiene un sistema de recomendación en tiempo real y sistemas de filtrado colaborativo sincronizado a las necesidades del investigador, lo que le ayuda a navegar y descubrir el trabajo de otros en su campo, nuevos descubrimientos de la investigación y los pesos del valor de la información presentada en la literatura de diferentes fuentes(Arévalo and Vázquez, 2016).

Debido a que las ALM están disponibles poco después de la publicación y se actualizan continuamente, estas métricas proporcionan una instantánea del alcance de un artículo en un momento dado. *PLoS ALM* extrae la información a contabilizar de las siguientes fuentes:

Visto : PLOS Journals (HTML, PDF, XML), PubMed Central (HTML, PDF)

Guardado : CiteULike, Mendeley

Citado : CrossRef, Datacitep, Europe PMC, PubMed Central, Scopus, Web of Science

[39] http://article-level-metrics.plos.org/

Recomendado : F1000 Prime

Discutido : Comentarios PLoS, Facebook, Reddit, Twitter, Wikipedia

2.4.7.10. Figshare

Figshare [40] es un repositorio de materiales de investigación multidisciplinar que fue fundado por Mark Hahnel en 2011; que posteriormente ha sido sostenido por Digital Science of Macmillan Publishers. *Figshare* ofrece una cantidad limitada de espacio de almacenamiento gratuito para uso privado (1 GB de almacenamiento gratuito ampliable mediante una cuota mensual), y una cantidad ilimitada de espacio de almacenamiento para materiales compartidos públicamente. Actualmente, los usuarios pueden subir figuras, multimedia (vídeos, audios), posters, publicaciones, tesis, código (código fuente y archivos binarios), presentaciones, bases de datos, y conjuntos de archivos (Kraker et al., 2015). *Figshare* es un repositorio donde los usuarios pueden hacer que todos sus productos de investigación estén a disposición de una manera citable, compartible y visible. Permite a los usuarios subir cualquier formato de archivo para ser previsualizado en el navegador de modo que cualquier resultado de investigación, desde posters y presentaciones a los conjuntos de datos y código, se difunda de una manera tal que el modelo actual de publicación académica no lo permite(Figshare, 2016).

A todo el material publicado a través de *Figshare* se le asigna automáticamente un DOI de *DataCite*, a través de la *California Digital Library* (de manera de animar a los usuarios a dar crédito a los autores de los recursos utilizados citándolo en publicaciones formales) y una referencia bibliográfica que enlaza directamente a su contenido en *Figshare*. También proporciona licencias *Creative Commons CC-BY* para las figuras, material audiovisual, posters, papers y grupos de archivos; y CC0 para bases de datos. Además de generar documentos QR de todo el material y otros mecanismos de integración y difusión en las principales redes sociales y gestores de referencias. La importancia de *Figshare* es que facilita la indización de sus contenidos por parte de los principales motores de búsqueda y bases de datos.

Figshare proporciona también estadísticas sobre las visualizaciones de tu documento incluyendo la cantidad de citas realizadas sobre este

[40] https://figshare.com/

por parte de otros documentos, Webs o medios. Por lo que es tenido en cuenta por parte de las principales herramientas altmétricas como *Almetrics.com*, *ImpactStory* o *Plum*; y por lo tanto una herramienta muy útil para la medición del impacto social de la investigación en asociación con otras herramientas como son *F1000 Research*, *PLOS ONE*, *Taylor & Francis* y *IOP Publishing* (Arévalo and Vázquez, 2016).

Con el fin de cumplir con los requisitos de metadatos *DataCite*, los usuarios están obligados a agregar la siguiente información antes de hacer públicos y citables los archivos: título, lista de autores, categorización de la disciplina, palabras clave y una descripción con tanto contexto como sea necesario para interpretar los archivos. Los autores también pueden añadir enlaces a fuentes externas. La estructura de los datos en *Figshare* es mínima: categorías (o disciplinas) son un conjunto cerrado de ontologías con 13 opciones. Las palabras clave son libres, pero se pueden autosugerir si etiquetas similares han sido utilizados por otros autores. Las descripciones del contenido son también a texto libre. Es importante señalar que todos los metadatos del contenido es generado por el autor. Una indización precisa se basa en el uso correcto de estos campos por parte del autor y en la definición adecuada del contenido en *Figshare*, la cual luego será indexado por *Google*.

La principal ventaja que posee *Figshare* frente a otros repositorios es que no está centrado en una disciplina específica, permite subir múltiples tipos de recursos y parece ser el actual ejemplo principal de este tipo de repositorios científicos universal (Thelwall and Kousha, 2016).

2.4.7.11. Bookmetrix

Bookmetrix[41] reúne un conjunto de métricas de rendimiento, que ayuda ver a un autor cómo sus libros se están discutiendo, citando, y utilizando en todo el mundo. *Bookmetrix* es una plataforma nueva y única de acceso gratuito desarrollada por *Springer* en asociación con *Altmetrics* que ofrece los datos de impacto que tienen los libros de *SpringerLink* [42] tales como el número de citas, descargas, lectores, reseñas de libros, entradas en la Wikipedia, redes sociales, etc. tanto a nivel de libro como de capítulo. Todo lo cual facilita una visión global del uso y alcance del libro (Springer, 2015).

[41] http://www.bookmetrix.com/
[42] http://link.springer.com/

Los datos capturados a través de *Bookmetrix* se muestran en las páginas del libro de *Springer* contenido en la plataforma *SpringerLink* e informa con qué frecuencia un libro o capítulo individual es mencionado, compartido, comentado o leído en línea. Los datos son actualizados en tiempo real, con lo que se pretende proporcionar una representación precisa del alcance actual, el uso y el impacto en general de cada libro o capítulo para todos los autores, editores y lectores.

Bookmetrix presenta estas métricas a través de una página general detallada sobre el libro (Bayaz, 2015), compuesta por 5 fichas o pestañas (como se observa en la Figura 2.17):

- La pestaña *Citas* muestra el número de citas, tanto a nivel del libro como del capítulo, basado en los datos recogidos por *CrossRef*.
- La pestaña *Menciones* utiliza los datos proporcionados por *Altmetrics* para mostrar a los usuarios cómo el libro/capítulo está siendo discutido, mencionado o compartido en medios en línea, incluyendo documentos políticos públicos, canales principales de noticias, blogs y una variedad de redes sociales.
- La pestaña *Lectores* ofrece información sobre cuántas personas han guardado el libro / capítulo en su administrador de referencias, incluyendo su país de origen y ocupación.
- La pestaña *Comentarios* muestra extractos de reseñas de libros conocidos por *Springer*.
- La pestaña *Descargas* muestra tanto los datos mensuales y el total de descargas del libro/capítulo como se registra a través de *SpringerLink*.

Bookmetrix no sólo le ofrece una visión global del uso y alcance de un libro, los datos proporcionados también ayudan a (Springer, 2015):

- contestar las preguntas de su proveedor de fondos a cerca del impacto general de su trabajo
- demostrar el compromiso que rodea su investigación
- descubren nuevos públicos y colaboradores potenciales

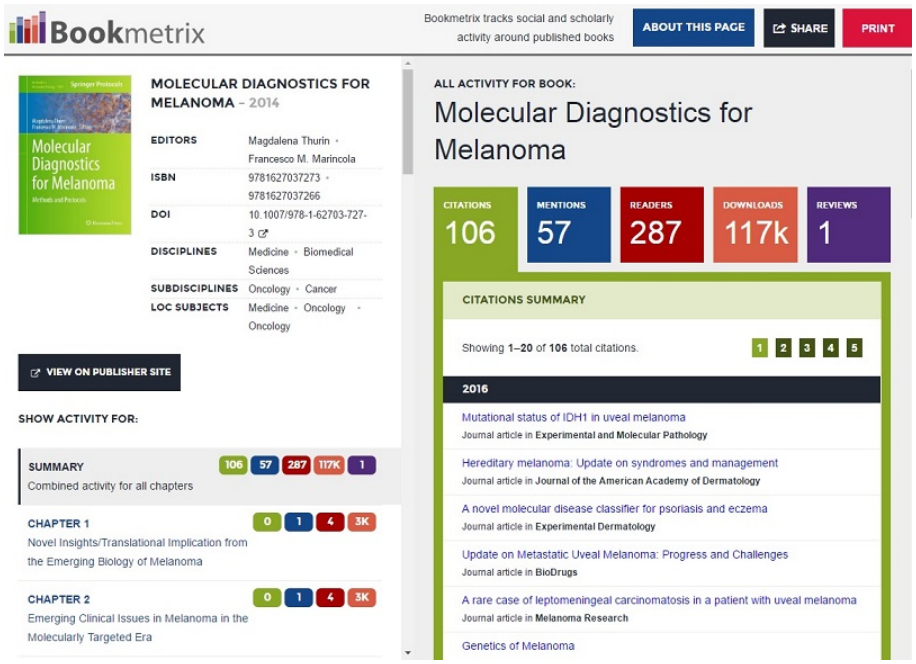


Figura 2.17: Ejemplo de Métricas proporcionadas para un libro desde el sitio web de *Bookmetrix*

- averiguar donde su trabajo recibe la mayor atención y de allí establecer relaciones continuas con los bloggers principales, responsables políticos o medios de comunicación
- comparar su trabajo frente a otros similares o compañeros en su campo de trabajo

2.4.7.12. Facebook

Facebook[43] es la red social más grande hoy en día y objeto de estudio desde distintos puntos de vista (en este trabajo sólo se hará mención al interés bibliométrico de esta plataforma). Según los datos que puedan ser de utilidad, las herramientas que se utilicen, recogerán datos de los perfiles públicos de usuarios, ya que es el objetivo de las *altmetrics*. En este sentido, como lo indica (AltmetricSupport, 2016) solo resulta de interés

[43] https://www.facebook.com/

los "posts" en páginas públicas (no así los posts en las líneas de tiempo individuales y los *likes*), en otros casos se pueden rastrear comentarios, las veces que se indica con "me gusta" un post o un comentario. A diferencia de las redes sociales y sitios web mencionados anteriormente, *Facebook* no es una red social académica, pero si una red social masiva, pública y de alcance mundial, con lo cual resulta más que atractiva a los investigadores a la hora de publicar sus artículos y avances, en dicha plataforma.

2.4.7.13. Twitter

Twitter[44] es la red de microblog más importante en la actualidad (al igual que con *Facebook*, en este trabajo sólo se hará mención al interés bibliométrico de ésta plataforma). La audiencia de esta plataforma crece a pasos agigantados, y la popularidad y el uso de *Twitter* como fuente de datos resulta más que interesante para innumerables análisis, estudios y comparaciones, tal es el caso de las *altmetrics*, como lo indica (Altmetric-Support, 2016), los datos que recolectan son solamente los comentarios públicos y *retweets* (no así los favoritos). El avance de los indicadores obtenidos con esta plataforma se debe principalmente al esfuerzo colectivo de desarrolladores y a la existencia de múltiples herramientas que hacen uso de la API para obtener datos de interés que puedan reflejar algún conocimiento relevante.

2.5. Open Content

Hasta no hace muchos años el acceso a la información era una restricción real y cotidiana, no solo hablando de los permisos o credenciales para acceder a un contenido digital, sino también, haciendo referencia a la brecha digital que separa a las personas de distintos lugares del mundo. Internet ha logrado disminuir esta diferencia, pero aun hoy, no todo el mundo está conectado y quienes están conectados no siempre son capaces de acceder a la información que necesitan, a causa de regulaciones del contenido, bloqueos a ciertos sitios que no respetan la ideología política del país, censura indiscriminada, entre tantos otros motivos.

[44] https://twitter.com/

La sociedad de la información se ha convertido en los últimos años en la sociedad de la información abierta. Como lo indica (Méndez, 2015)... "Casi todos los conceptos de moda en nuestro entorno profesional incluyen, de una u otra manera la palabra abierto (open, en inglés, que tiene más glamour): open access, open content, open data, open research, open education, open innovation, open knowledge, etc.". Si bien no existe un consenso claro sobre lo que se considera conocimiento abierto, siguiendo la definición de *opendefinition.org*: "El conocimiento es abierto si cualquiera es libre para acceder a él, usarlo, modificarlo y compartirlo bajo condiciones que, como mucho, preserven su autoría y su apertura."(Dutta, 2015; OpenDefinition, 2016)

El *Open Access (OA)* es un tipo de acceso, no es un tipo de modelo de negocios, de licencia o de contenidos. Las obras disponibles en *OA* son digitales, gratis, están en línea y mayormente libres de derechos de autor y otras restricciones de uso(Suber, 2012). Una definición más profunda de *OA* es la que ofrece la *Budapest Open Access Initiative* [45]: "Por acceso abierto a estas obras se entiende que están disponibles gratuitamente en Internet y que se permite que cualquier usuario lea, descargue, copie, distribuya, imprima, busque y enlace el texto completo de estos artículos, que se los rastree para indexarlos, que se los transforme en datos para ser incluidos en programas informáticos y se los use con cualquier otro propósito legítimo, sin que medie ningún impedimento financiero, legal o técnico. En este sentido, el único rol de los derechos de autor debe ser el de darle a los autores el control sobre la integridad de su obra así como el derecho de ser reconocidos y citados apropiadamente".

La iniciativa *Open Contents* se refiere a la distribución, uso, copia y modificación gratuita de los resultados de cualquier actividad creativa. Esto incluye una amplia gama de recursos, pero ha tenido un impacto más profundo en una serie de áreas, muchas de ellas relacionadas con las actividades de las instituciones de educación superior, tanto en los recursos educativos y culturales como en las actividades de investigación científica (Tomlin, 2009; García-Peñalvo et al., 2010).

En otras palabras, "Abierto" se refiere al hecho de otorgar permisos de *copyright* más allá de los ofrecidos por la ley de *copyright* estándar. Desde un punto de vista quizás simplista, pero intuitivo, cuanto menos restricciones se impongan a cierto contenido, más "abierto" resulta ese

[45] http://www.budapestopenaccessinitiative.org/read

contenido (García-Peñalvo et al., 2010). En este sentido, (Suber, 2008) habla de "gratis Open Access" para la eliminación de las barreras de precios por sí sola y "libre Open Access"para la eliminación de precios y al menos algunas barreras de permisos. Los nuevos términos permiten hablar sin ambigüedad acerca de estos dos tipos de acceso gratuitos en línea.

En consecuencia, los términos *Open Content - OC (Contenido abierto)* y *Open Educational Resources - OER (Recursos Educativos Abiertos)* describen cualquier trabajo con *copyright* (tradicionalmente el software es excluido, ya que se describe bajo otros términos como *open source*) que posee licencia de modo de proporcionar a los usuarios acceso libre y perpetuo para realizar las siguientes cinco actividades (Wiley, 2014; OpenContent.org, 2016):

1. Conservar: el derecho de hacer, poseer, y controlar ejemplares del contenido (por ejemplo, descargar, duplicar, almacenar y gestionar)

2. Reutilizar: el derecho a utilizar el contenido en una amplia gama de formas (por ejemplo, en una clase, en un grupo de estudio, en un sitio web, en un video)

3. Revisar: el derecho de adaptar, ajustar, modificar o alterar el contenido mismo (por ejemplo, traducir el contenido a otro idioma)

4. Combinar: el derecho de combinar el contenido original o revisado con otros materiales para crear algo nuevo (por ejemplo, incorporar el contenido a un *mashup*)

5. Redistribuir: el derecho de compartir copias del contenido original, sus revisiones, o sus combinaciones con los demás (por ejemplo, dar una copia del contenido a un amigo)

Los repositorios que adoptan este esquema son de libre acceso, lo cual significa que los motores de búsqueda y rastreadores de información pueden indizar el contenido de estos, permitiendo el acceso no sólo a los metadatos como es el caso de la iniciativa *OAI-PMH* (Ginsparg et al., 1999; Van de Sompel and Lagoze, 2000), sino también al contenido completo del material digital. La OAI y el OA son iniciativas separadas que sin embargo se solapan, pero no deben ser confundidas una con otra.

Una comparación entre *Google Scholar* y *Scopus* propuesta por (Moed et al., 2016) puede dar un indicio de los sesgos que poseen las bases de datos tradicionales (las revistas indexadas por este tipo de base de datos eligen qué artículos publicar, los autores deben pagar para que su artículo sea publicado, no todas las revistas son indexadas) frente a *Google Scholar* (que indexa todo lo que puede y considera como contenido académico). Esto se evidencia con las revistas de acceso abierto que no son incluidas en las bases de datos tradicionales, principalmente por no ser consideradas de alto impacto, al contrario, *Google Scholar* si las incorpora y son estas revistas la principal fuente de datos de este motor.

2.6. Desambiguación

La desambiguación de nombres es un tema de profundo interés dentro de la Recuperación de Información y en todo lo que respecta al análisis bibliográfico, ha sido estudiado desde distintos puntos de vista y aún no posee una respuesta que satisfaga a todas las variantes y demandas. Existen muchas y variadas soluciones, y un número grande de acepciones de este mismo problema (Han et al., 2004), como pueden ser: vinculación de registros (Fellegi and Sunter, 1969), fusión/purga (Hernández and Stolfo, 1998), detección de duplicados (Bitton and DeWitt, 1983; Elmagarmid et al., 2007; Lee et al., 2000; Low et al., 2001; Monge and Elkan, 1997), endurecimiento de bases de datos (Cohen et al., 2000), incertidumbre de identidad (Pasula et al., 2003), resolución de correferencia (Soon et al., 2001), y coincidencia de nombres (Bilenko et al., 2003; Branting, 2002; Tejada et al., 2002; Valarakos et al., 2004). Esta diversidad se ha visto reflejada en las investigaciones de las siguientes áreas: estadísticas, bases de datos, bibliotecas digitales, procesamiento del lenguaje natural y minería de datos.

Como indica (Kern et al., 2011), la ambigüedad puede tener como origen algunas de las siguientes causas:

1. Dos autores distintos que comparten la misma representación del nombre de autor, por ejemplo: Adam Smith.
2. Un mismo autor es referenciado por diferentes variaciones ortográficas, por ejemplo: Jose Juarez por Jose Juares.
3. Un autor determinado ha cambiado su nombre a través del tiempo.

Además de estas causas, existen otras que tienen como origen errores humanos, tales como:

1. Errores de escritura, traducción o abreviaturas en los nombres.
2. Cuando un autor firma de forma diferente los trabajos, por ejemplo: en algunos trabajos utiliza ambos nombres y en otros solo uno.
3. Las diferentes formas de indexado de las base de datos bibliográficas.

La ambigüedad de nombres puede afectar a la calidad de la recopilación de datos científicos, la integridad de las bases de datos, puede disminuir el rendimiento de la recuperación de información y búsqueda en la Web, y puede incluso causar la incorrecta identificación y atribución de crédito a los autores(Han et al., 2005, 2004).

Si bien la mayoría de los estudios se han ocupado en resolver la ambigüedad relativa a los nombres de los autores, es necesario destacar que de la misma forma se pueden presentar ambigüedades en los títulos de las publicaciones o en otros campos de datos como la dirección postal, por ello en este trabajo el esquema utilizado resuelve el problema de la desambiguación tanto en los nombres de los autores como en los títulos de las publicaciones, ya que como lo expone (Beel and Gipp, 2010) al recuperar los registros de los motores de búsqueda se pueden obtener publicaciones repetidas (dos registros que hacen referencia a la misma publicación, web spam), o errores u omisiones en el título de estas.

El esquema que se decida emplear para resolver el problema de la desambiguación dependerá de varios factores: la naturaleza del problema, el origen y tipo de datos, la cantidad de información, el o los campos que se deseen desambiguar (no es lo mismo romper la ambigüedad en campos como las cuentas de email que hacerlo con los nombres de personas o direcciones postales), entre otros. De todos modos la variante más estudiada ha sido la desambiguación de nombres de autores, debido a que este campo es uno de los campos más sensibles a la falta de normalización ya que las variaciones sobre la forma en que figura un autor en sus diferentes publicaciones pueden dificultar el cálculo correcto de su productividad (Costas and Bordons, 2007).

Dicha variante, como lo indica (Lee et al., 2005), posee dos orígenes bien diferenciados. El primero de ellos denominado *Mixed Citation (MC)*

se presenta cuando dos o más autores poseen exactamente el mismo nombre, son homónimos, con lo cual los datos de las citas se fusionarán erróneamente en un único autor derivando en un análisis incorrecto de las mismas. El segundo, *Split Citation (SC)* o también más conocido como *Name matching*(Han et al., 2005), ocurre cuando en una misma biblioteca o base de datos, un autor posee diferentes variantes o etiquetas para su nombre cuando de hecho se refieren al mismo autor, por ejemplo, para un autor llamado "Ji-Woo K. Li", algunos posibles errores pueden ser las abreviaturas ("J. K. Li"), alternación de nombres ("Li, Ji-Woo K."), de tipo ("Ji-Woo K. Lee" o "Jee-Woo K. Lee"), contracciones ("Jiwoo K. Li"), omisiones ("Ji-Woo Li") o combinaciones de estas.

2.6.1. Comparaciones exactas vs comparaciones aproximadas

En relación al problema mencionado anteriormente (*Split Citation*), los motores abiertos o de libre acceso presentan otro problema importante. Entre los resultados de una búsqueda determinada, además de existir distintas variantes para un autor dado, entre los resultados se encuentran registros que no pertenecen al criterio seleccionado, no solo por una mala atribución, sino, en el caso específico de *Google Scholar*, este motor también recoge registros con parte del nombre del autor especificado. Mientras la búsqueda sea más especifica (posea más términos, sin abreviaturas) los resultados estarán mejor ajustados, por el contrario (buscar autores solo por un apellido o incluyendo iniciales) deriva en resultados poco precisos y en una tarea mayor de depuración y limpieza.

Como se mencionó, algunos resultados de *Google Scholar* no se encuentran normalizados, entre estos se pueden encontrar resultados duplicados, resultados que no pertenecen al autor solicitado, por ejemplo si la búsqueda se realiza por "Jose Luis Ortega" entre los resultados existen trabajos de "Jose Luis Ortega -Priego" o "Jose Luis Gallego Ortega", resultados donde parte de los términos de búsqueda se hallan en el título o resumen de la publicación o en alguna parte del cuerpo de la publicación (lo más común son las citas bibliográficas al final de los trabajos). Para la misma búsqueda, entre las variantes de nombres correctas se encuentran: JL Ortega, Jose Luis Ortega, J Luis Ortega, OJ Luis (Medrano et al., 2012a).

En este tipo de problemas las búsquedas exactas no resultan eficien-

tes debido a que las variaciones entre un registro de dato y otro muchas veces son pequeñas (algunos caracteres debido a errores u omisiones) o se producen alternación de términos o abreviaturas, por ello el concepto de "near duplicate" (Yang and Callan, 2006; Xiao et al., 2011; Rico-Sulayes, 2015) o "duplicado aproximado-próximo-cercano" resulta relevante, en este sentido dos registros serán duplicados aproximados si son *por lo menos idénticos*. Esto se calcula con alguna de las funciones de distancia de edición que se verán a continuación, estas funciones entregan como resultado un número que indica cuan *idénticas* son dos cadenas de caracteres, y al sobrepasar un umbral definido por el usuario son consideradas idénticas.

2.6.2. Técnicas de desambiguación

El problema de la desambiguación posee muchas variantes, algunas más comunes y estudiadas que otras, y las soluciones aportadas en la materia resuelven o intentan resolver las distintas generalizaciones del problema. Como se mencionó anteriormente las causas de este problema son múltiples, y la forma más común del problema se podría definir de la siguiente manera (On et al., 2005):

> Dadas dos listas *largas* de nombres de autores, X and Y,
> para cada nombre de autor x $(\in X)$, encontrar un conjunto de
> nombres de autores $y_1, y_2, ..., y_n$ $(\in Y)$ tal que ambos
> x y y_i $(1 \leq i \leq n)$ sean variantes del nombre
> del mismo autor.

De la misma forma (On et al., 2005) indica que el enfoque base para resolver el problema es tratar a cada nombre de autor como una "cadena de caracteres", y ejecutar para cada uno de los pares de nombres alguna función de distancia entre estas, $dist(x, y)$:

```
for each name x(∈ X)
    for each name y(∈ Y)
        if dist(x, y) > φ, x and y are name variants
```

Esta aproximación es bastante básica y resulta prohibitiva para el

caso de listas de nombres muy grandes (porque posee un tiempo de resolución de complejidad cuadrática $O(|X||Y|)$), por ello es necesario implementar algoritmos escalables que no dependan de las similaridades sintácticas de los nombres de autores.

Uno de los esquemas que mejor resultado proporcionan es la división del conjunto de nombres a evaluar en conjuntos (sets o bloques), esta operación es normalmente conocida como "blocking" o "clustering". La idea central es dividir el conjunto de nombres en conjuntos más pequeños y luego realizar la evaluación de la distancia entre bloques y no entre todos los nombres. La función de asignación de cada nombre a un conjunto determinado tiene especial importancia, ya que depende de ésta por un lado el tamaño de los conjuntos y la exactitud de los elementos dentro de cada conjunto.

El *blocking* o *clustering*, al igual que otros mecanismos de desambiguación pueden ser implementados utilizando métodos de aprendizaje supervisado (Han et al., 2004; Bilenko and Mooney, 2003; Seol et al., 2016; Lerchenmueller and Sorenson, 2016; Zhang et al., 2016), no supervisado (Han et al., 2005; Momeni and Mayr, 2016) o en combinación de ambos esquemas (Kang et al., 2009; On et al., 2005), los primeros requieren un entrenamiento de la función de asignación (en el caso del blocking) para establecer los criterios con los cuales una entrada (nombre de autor) es asignada a un bloque o cluster, en este caso se seleccionan un conjunto de entradas representativas de una población determinada y el algoritmo es entrenado para afinar los parámetros de selección. Estos mecanismos por lo general requieren una mayor cantidad de tiempo inicial (para el entrenamiento) pero redundan en una mayor precisión y se utilizan cuando se cuenta con bases de datos no muy grandes, ya que para grandes bases de datos resulta inapropiado por el alto costo computacional(Huang et al., 2006). El segundo tipo de métodos, los no supervisados, se utilizan generalmente cuando no se conocen los datos iniciales a desambiguar o no es posible contar con ellos, por ejemplo en el caso de búsquedas en tiempo real o con grandes bases de datos o cuando la cantidad de datos a desambiguar es muy pequeña.

Otros esquemas en cambio, aprovechan las ventajas de las Redes Neuronales Artificiales, como es el caso de (Tran et al., 2014) que utilizó un esquema basado en *Deep Neural Network* para aprender automáticamente características para resolver el problema de la ambigüedad en nombres de autores, adicionalmente propone la arquitectura general del sistema

para la desambiguación de nombres sobre cualquier conjunto de datos.

En (Ferreira et al., 2012) se ofrece una taxonomía y un resumen de los métodos, mas importantes y relevantes, utilizados para resolver este problema.

2.6.3. Funciones de distancia de edición

Algunos autores como (Bilenko et al., 2003; On et al., 2005), resuelven el problema de la desambiguación mediante funciones que evalúan la similaridad entre dos cadenas de caracteres para detectar los duplicados. En este tipo de funciones se evalúan las distancias de edición, definida como: la distancia de edición entre una cadena de caracteres s y t es el coste de la mejor secuencia de operaciones de edición (inserción, borrado y sustitución) de caracteres individuales para convertir s en t(Levenshtein, 1966). Esta distancia de edición es comúnmente conocida en la bibliografía como distancia de *Levenshtein*, en honor a su creador, existen otras distancias como *Damerau-Levenshtein* (Damerau, 1964; Levenshtein, 1966) o *Needleman-Wunsch* (Needleman and Wunsch, 1970), pero son generalizaciones de la primera.

El esquema propuesto por *Levenshtein* es muy conocido, y de fácil aplicación (Ristad and Yianilos, 1998; Serva and Petroni, 2007; Snae, 2007; Wichmann and Holman, 2009; Tria et al., 2010; Greenhill, 2011; Chacón et al., 2014; Abdulkhudhur et al., 2016). Por ejemplo la distancia de *Levenshtein* entre *casa* y *caza* es de 1, ya que se sustituye la s por la z. Esta medida normalmente es normalizada dividiendo el resultado por la longitud de la palabra más grande (Serva and Petroni, 2007), con lo cual la distancia entre *casa/caza* es 0.25.

Una función un tanto más compleja que la anterior, es la función de distancia *Monge-Elkan* (Monge and Elkan, 1996), la cual es una variante de la función de distancia *Smith-Waterman* (Durbin et al., 1999).

Una métrica muy similar es la métrica de *Jaro* (Jaro, 1989, 1995; Winkler, 1999) la cual se basa en el número y orden de los caracteres comunes entre dos cadenas de caracteres. Dadas las cadenas de caracteres $s = a_1...a_K$ y $t = b_1...b_L$, definir un caracter a_i dentro de s para tener "en común" con t sí y solo si existe un $b_j = a_i$ dentro de t tal que $i - H \leq j \leq i + H$, donde $H = min(|s|, |t|)/2$. Permitiendo que $s' = a'_1...a'_K$ sean caracteres dentro de s que son en común con t (en el mismo orden

que aparecen en s), y permitiendo que $t' = b'_1...b'_L$ sean los análogos. A continuación definir una transposición de s', t' para ser una posición i tal que $a_i = bi$. Permitiendo que $T_{s',t'}$ sea la mitad del número de transposiciones de s' y t'. La métrica de Jaro para s y t es:

$$Jaro(s,t) = \frac{1}{3}.\left(\frac{|s'|}{|s|} + \frac{|t'|}{|t|} + \frac{|s'| - T_{s',t'}}{s'}\right)$$

William Winkler (Winkler, 1999) propuso una variante de la métrica de Jaro, que además utiliza la longitud P del prefijo en común más largo de s y t. Siendo $P' = max(P, 4)$, se define:

$$Jaro - Winkler(s,t) = Jaro(s,t) + \frac{P'}{10}.(1 - Jaro(s,t))$$

Las métricas de Jaro y Jaro-Winkler parecen estar destinadas principalmente para la comparación de cadenas de caracteres cortas, como los nombres o apellidos.

2.6.4. Funciones de distancia basadas en token (*token-based*)

En algunas situaciones, el orden de las palabras es insignificante. Por ejemplo, las cadenas de caracteres "Wong Lee" y "Lee, Wong" es probable que sean duplicados, incluso si no son cercanos en la distancia de edición. En estos casos de podrían convertir las cadenas de caracteres s y t a conjuntos múltiples de tokens(donde cada palabra es una muestra) y considerar las métricas de similaridad en estos conjuntos.

Una métrica basada en tokens es la *similitud de Jaccard* (Braam et al., 1988; Hamers et al., 1989; Chaudhuri et al., 2006). La similitud de Jaccard entre los conjuntos de palabras S y T se define como:

$$similitud\ de\ Jaccard(S,T) = \frac{|S \cap T|}{S \cup T}$$

Una métrica de uso amplio por la comunidad de recuperación de información es la *frecuencia del término - inversa de la frecuencia del*

documento (del nombre y siglas en inglés: *term frequency-inverse document frequency (TF-IDF)*) o también conocida por *similitud del coseno (cosine similarity)*(Aizawa, 2003; Hiemstra, 2000; Philbin et al., 2007; Tata and Patel, 2007; Chum et al., 2008), definida de la siguiente forma:

$$TF - IDF(S,T) = \sum_{w \in S \cap T} V(w,S).V(w,T)$$

Donde $TF_{w,S}$ es la frecuencia de la palabra w en S, IDF_w es la inversa de la fracción de los nombres dentro del cuerpo del texto que contienen w,

$$V'(w,S) = log(TF_{w,S} + 1).log(IDF_w)$$

,y

$$V(w,S) = \frac{V'(w,S)}{\sqrt{\sum_{w'} V'(w,S)^2}}$$

Para proceder con el emparejado de nombres, se pueden recolectar las estadísticas usadas para calcular IDF_w del conjunto completo de nombres que se utilizarán para comprobar las coincidencias.

En Dagan *et al* (Dagan et al., 1999), un conjunto de tokens S puede verse como muestras de una distribución desconocida de tokens P_S, y la distancia entre S y T puede calcularse basándose en dicha distribución. Para ello se considera la distancia *Jensen-Shannon* (Melville et al., 2005) entre P_S y P_T. Siendo $KL(P||Q)$ la divergencia *Kullback-Lieber* (Kullback and Leibler, 1951) y siendo $Q(w) = \frac{1}{2}(KL(P_S||Q) + KL(P_T||Q))$,

$$Jensen - Shannon(S,T) = \frac{1}{2}(KL(P_S||Q) + KL(P_T||Q))$$

Un método propuesto por Fellegi y Sunter en el área de vinculación de registros (Fellegi and Sunter, 1969), pude ser fácilmente extendido como una función basada en la distancia de tokens. Siendo A y B dos conjuntos de registros a verificar sus coincidencias, siendo $C = A \cap B$, $D = A \cup B$, y para $X = A, B, C, D$ permitiendo que $P_X(w)$ sea la probabilidad empírica de un nombre que contiene la palabra w dentro

del conjunto X. También, siendo e_X la probabilidad empírica de un error en un nombre en el conjunto X; $e_{X,0}$ la probabilidad de la ausencia de un nombre dentro del conjunto X; e_T la probabilidad de dos nombres correctos pero diferentes en A y B; y siendo $e = e_A + e_B + e_T + e_{A,0} + e_{B,0}$. Fellegi y Sunter proponen pares de ranking s, t por medio del índice de probabilidad:

$log(\frac{P_r(M|s,t)}{P_r(U|s,t)})$

donde M es la clase de pares coincidentes y U es la clase de pares no coincidentes. Bajo una plausible serie de supuestos, se puede aproximar el resultado incremental para el índice de probabilidad asociado con el evento "s y t coinciden en contener la palabra w" utilizando $log(IDF_w)$.

2.6.5. Esquemas híbridos

Monge y *Elkan* proponen el siguiente esquema recursivo para la equiparación de dos cadenas de caracteres, s y t, de gran tamaño. Primero, s y t se dividen en subcadenas de caracteres $s = a_1...a_K$ y $t = b_1...b_L$. De este modo, la función de similaridad quedaría del siguiente modo:

$$sim(s,t) = \frac{1}{K}\sum_{i=1}^{K} max_{j=1}^{L} sim'(A_i, B_j)$$

donde sim' es alguna función de distancia secundaria que logra buenos resultados trabajando sobre cadenas de caracteres pequeñas (tal como Jaro-Winkler). El método Monge-Elkan ofrece un mecanismo para medir la similitud entre dos cadenas de caracteres que contienen unos pocos términos. Esta medida no es simétrica (Jimenez et al., 2009), con lo cual se debe prestar atención cuan necesario es este requerimiento pero sobre todo dependerá del problema en cuestión y del ámbito de aplicación. Por ejemplo, utilizando este método el nombre "Alvaro E. Monge" coincide con "A. E. Monge" mientras que a la inversa no es necesariamente cierto, pues "A. E. Monge" puede estar hablando de otra persona distinta (Monge, 2001).

También se puede considerar una versión "flexible"(soft) de TF-IDF, donde los tokens similares son considerados además tokens dentro de $S \cap T$. Nuevamente, siendo sim' una función de similaridad secundaria, y siendo $CLOSE(\theta, S, T)$ el conjunto de palabras $w \in S$ tal que existe

algún $v \in T$ tal que $dist'(w, v) > \theta$, y para $w \in CLOSE(\theta, S, T)$, siendo $D(w, T) = max_{v \in T} dist(w, v)$. Se define:

$$SoftTF - IDF(S,T) = \sum_{w \in CLOSE(\theta,S,T)} V(w,S).V(w,T).D(w,T)$$

Para un mayor detalle de estás métricas revisar (Bilenko et al., 2003; Cohen et al., 2003).

2.6.6. Similitud fonética

Existen otras técnicas basadas no en las diferencias sintácticas o morfológicas de una palabra, sino en el sonido de estas. Podría resultar de utilidad encontrar las variantes de una palabra (un nombre, una materia, un lugar) que suene similar a otro. *Soundex* (Knuth, 1998) es un algoritmo que transforma un nombre en un código alfanumérico, esta clave generada está formada por una letra (la primer letra del nombre que se está procesando) seguido de un código de tres dígitos. Las claves generadas tienen la propiedad que palabras pronunciadas de forma similar producen la misma clave *soundex* y por lo tanto puede ser usada para simplificar búsquedas en bases de datos donde se conoce la pronunciación pero no la ortografía. *Soundex* fue desarrollado por Robert C. Russell y Margaret K. Odell y patentado en 1918 y 1922 (Russell, 1918).

Esta técnica ha sido utilizada como método de desambiguación al comparar dos cadenas de una forma mucho mas sencilla y rápida, y en conjunto con otros esquemas basados en *tokens* (Tang and Walsh, 2010; Kalmar and Freitag, 2009).

Metaphone (Philips, 1990) es un algoritmo fonético similar a *Soundex*, utilizado para indexar palabras por su sonido al ser pronunciadas en Inglés. Es más exacto que *Soundex* y genera claves de longitud variable. Mas adelante, el autor desarrolló *Double Metaphone* (Philips, 2000) que es una mejora del algoritmo *Metaphone* anterior. Ambos algoritmos se utilizan para devolver una aproximación de cómo suenan las palabras en Inglés, la clave retornada debe ser la misma para las palabras o nombres que suenan similares, y se puede utilizar como clave de búsqueda. Los algoritmos *Metaphone* son ampliamente utilizados en correctores or-

tográficos, interfaces de búsquedas, programas de autentificación y en búsquedas de genealogía.

2.6.7. Métodos basados en blocking/clustering

El estudio de ((Han et al., 2005)) propone un enfoque de aprendizaje no supervisado utilizando el método de *clustering* o agrupamiento *K-way spectral clustering* para desambiguar los nombres de los autores en las citas bibliográficas. Dicho enfoque utiliza tres tipos de atributos de las citaciones: los nombres de los co-autores, los títulos de las publicaciones y el origen de las publicaciones (como ser el nombre de la revista o congreso). Para ello se utilizaron conjuntos de datos extraídos de la base de datos bibliográfica DBLP. Han *et.al.* demuestran mediante diferentes variantes de este enfoque (variando el tamaño de los conjuntos de datos, la diversidad de las áreas de investigación de los autores, entre otros) cómo se puede ver afectada el desempeño de la desambiguación. También muestra, que mediante el empleo de características adicionales (los nombres de los co-autores, los títulos de las publicaciones y el origen de las publicaciones) la precisión en la desambiguación mejora notablemente.

Por su parte (On et al., 2005) emplea un mecanismo basado en dos pasos, el primer paso es reducir la cantidad del número de candidatos a comparar mediante *blocking* y el segundo paso es medir la distancia de edición de dos nombres por medio de la información de los coautores.

Otro estudio que utiliza la información de los coautores y variables colaborativas para resolver este problema es (Li et al., 2014), quien utilizando un esquema iterativo de *blocking* y *clustering*, y un clasificador basado en el teorema de Bayes (Lewis, 1998; Bishop, 2006), logró aplicarlo a las bases de datos de patentes de los Estados Unidos de Norteamérica entre los años 1975-2010. Un estudio similar, que también emplea la información de la red de coautores es el que presenta (Seol et al., 2016), el cual utiliza un esquema supervisado de *Support Vector Machine (SVM)* en combinación con la información relacionada del autor, en este caso: e-mail, afiliación, campo de estudio principal y palabras claves. Los resultados demostraron que cuando esta información se combina a la red de coautores se logran resultados favorables.

2.6.8. Iniciativas basadas en un identificador único

Al margen de las soluciones, técnicas, y herramientas enumeradas anteriormente, en la actualidad existen iniciativas más que válidas y en crecimiento para resolver el problema de la ambigüedad entre las publicaciones de los autores y entre los autores. La idea de estos esquemas es resolver parte del problema, principalmente el relacionado con la atribución errónea de créditos a un autor determinado. Por ejemplo, para el caso de que un autor publicara un Artículo A, y la base de datos bibliográfica lo indexara de dos formas distintas como artículo A_1 y artículo A_2, (el origen de esta doble indexación y con lo cual la existencia de un duplicado, puede deberse a que el motor de indexación tomó de dos sitios distintos el mismo artículo o simplemente lo almacenó como proveniente de dos revistas distintas, por un error ortográfico en el nombre de la revista, o por una variante en el nombre de la revista, como lo indica (Valderrama-Zurián et al., 2015) en su estudio, el cual encontró que para un registro duplicado, un registro era asignado a la revista *Psychoterapia* y el otro registro mapeado a *Archivesof Psychiatry and Psychotherapy*, ambas revistas publicadas por la Asociación Psiquiátrica de Polonia) luego este artículo, en sus dos versiones, recibe citas, el artículo A_1 recibe C_1 citas y el artículo A_2 recibe C_2 citas.

La existencia del registro duplicado existe pues las citas a ambos documentos pueden estar repetidas, esto quiere decir que la intersección entre C_1 y C_2 no es nulo. Si fuese posible hacer referencia de forma única a un artículo o documento bibliográfico, sin importar cómo es indexado, este problema no existiría, pues las citas y referencias siempre harían mención a un único nombre o dirección invariable. Esta es la aproximación que realiza el *Digital Object Identifier (DOI)* [46] para hacer referencia a un documento de forma unívoca e invariable a través del tiempo y sin importar dónde esté almacenado dicho documento.

En el mismo sentido, *Open Researcher and Contributor ID (ORCID)* [47] intenta resolver el problema de la existencia de variantes de nombres para un mismo autor, el origen de este problema es múltiple, puede deberse a una mala indexación del motor de base de datos (para un trabajo indexó al autor como Daniel Tores-Salinas y para otro como D. Torres-Salinas), o el autor firma de forma diferente los artículos con el paso del

[46] https://www.doi.org/
[47] http://orcid.org/

tiempo (inicialmente incluye en su firma todos los nombres y apellidos, y luego de un tiempo solo incluye el nombre de pila y los apellidos) para la base de datos son dos autores distintos y son tratados como dos individuos, siempre y cuando no exista una herramienta, mecanismo u opción que permita resolver este problema. Por lo tanto, si fuese posible hacer referencia a un autor mediante un código único e invariable y que no dependa del motor de indexación ni del repositorio dónde se encuentra almacenado el artículo, el problema estaría resuelto y no existiría un asignación errónea de créditos a un autor que no corresponde.

2.6.8.1. DOI

Digital Object Identifier (DOI), que puede traducirse como identificador digital de un objeto, es un estándar para la identificación única y permanente de contenidos en línea, y para la vinculación a través de Internet. Se trata de una cadena alfanumérica única, asignada y administrada por una agencia de registro (*International DOI Foundation - IDF*) [48]. Todas las cadenas DOI comienzan con un "10" (código de la agencia registradora del DOI, como solo es la IDF la encargada, este numero es fijo por el momento) y constan de un prefijo y un sufijo separados por una barra (esta conjunción es llamada nombre DOI). El prefijo es una combinación de la parte "10", que identifica el registro de DOI, seguido por varios (generalmente cuatro) caracteres alfanuméricos, que identifican a la agencia solicitante. En caso de que el solicitante del registro sea la misma IDF, la segunda parte del prefijo es "1000". El prefijo se asigna a una organización que desea registrar nombres DOI, cualquier organización puede optar por tener múltiples prefijos. Por otra parte, el sufijo identifica el objeto o contenido específico relacionado con este identificador y es elegido por el propio solicitante de registro. Son permitidos la mayoría de los caracteres *Unicode* legales, no haciendo diferencia entre letras mayúsculas o minúsculas (un ejemplo de DOI puede ser el siguiente: 10.1016/j.joi.2015.11.008).

Un nombre DOI es un identificador (no una ubicación) de una entidad en las redes digitales. Proporciona un sistema permanente y procesable para la identificación y el intercambio interoperable de información gestionada en las redes digitales. Un nombre DOI se puede asignar a cualquier entidad - física, digital o abstracta - sobre todo para compartir con

[48] https://www.doi.org/doi_handbook/7_IDF.html

una comunidad de usuarios interesados o para la gestión de la propiedad intelectual. El sistema DOI fue iniciado por la IDF (una organización sin fines de lucro, basada en miembros, fundada por varias asociaciones comerciales de publicación) en 1998, y más tarde estandarizado como ISO 26324.

Los nombres DOI existentes se pueden resolver de forma gratuita. El costo de registrar nuevos nombres DOI depende de los servicios utilizando un DOI provisto por una agencia de registro. Cada agencia de registro es libre de ofrecer su propio modelo de negocio en cumplimiento de las políticas generales de DOI. Agencias de registro individuales adoptan normas apropiadas para su comunidad y aplicación. Las agencias de registro son las encargadas de proveer los servicios del DOI a las entidades registradoras. En la actualidad las agencias de registro son: Airiti, Inc., Crossref, China National Knowledge Infrastructure (CNKI), Datacitep, EIDR (Entertainment Identifier Registry), ISTIC (The Institute of Scientific and Technical Information of China), JaLC (Japan Link Center), Korea Institute of Science and Technology Information (KISTI), mEDRA (Multilingual European DOI Registration Agency) y OP (Publications Office of the European Union).

El DOI es un identificador permanente para cualquier tipo de objeto o contenido (por ejemplo, textos, imágenes, tablas, registros de audio o vídeo, software, datos de investigación, etc.), tanto en las formas electrónicas y físicas, y puede referirse a diferentes niveles jerárquicos (por ejemplo, título de la revista, artículo, tabla o imagen incluidos en el artículo, libro o capítulo de un libro). Sin embargo, la estructura debe reflejarse en los metadatos de acuerdo con el *indecs Content Model*. *Indecs* es un acrónimo de "interoperability of data in e-commerce systems (interoperabilidad de los datos en los sistemas de comercio electrónico)" y fue un proyecto financiado por la Comunidad Europea con el fin de proporcionar un análisis de los requisitos de los metadatos del contenido del comercio electrónico.

Gran parte de la información de los metadatos, incluyendo la ubicación digital del objeto en Internet (tal como la URL), se almacena dentro del DOI. El DOI permanece inalterado durante toda la vida del objeto, mientras que los metadatos pueden cambiar. Por lo tanto, un DOI es una opción de vinculación más sólida que una simple referencia a un localizador uniforme de recursos (URL) que puede variar debido a la negligencia de mantenimiento de metadatos realizada por el editor.

El propósito principal del sistema DOI no sólo es la gestión de un conjunto de identificadores, sino también su capacidad de funcionamiento y la interoperabilidad. Las organizaciones que cumplan con los requisitos contractuales y están dispuestas a pagar una cuota de suscripción están autorizados para asignar los DOI (Gorraiz et al., 2016).

Según lo indica el DOI HandBook[49], el sistema DOI ofrece un conjunto único de funcionalidades:

- *Persistencia*, si el material se mueve, reorganiza o es marcado como favorito.
- *Interoperabilidad* con otros datos de otras fuentes.
- *Extensibilidad* mediante la adición de nuevas características y servicios a través de la gestión de los grupos de nombres DOI.
- *Gestión de datos única* para múltiples formatos de salida (independencia de la plataforma)
- *Gestión de la clase* de aplicaciones y servicios
- *Actualización dinámica* de metadatos, aplicaciones y servicios.

Los beneficios de implementar el sistema DOI incluyen facilitar la gestión de contenidos internos, permitiendo el desarrollo de productos más rápido, más escalable, haciéndolo más fácil y barato, por medio de las siguiente ventajas claves:

- Sabe lo que tiene (usuarios capaces de mirar catálogos de contenido disponibles en toda la empresa).
- Encuentra lo que buscas (usuarios capaces de buscar y navegar por el contenido a ser usados o re-utilizados).
- Saber dónde existe (capaz de ver donde se encuentra el elemento dentro de la organización)
- Ser capaz de conseguirlo (usuarios y herramientas de producción capaz de recuperar el contenido)

[49] https://www.doi.org/hb.html

El servicio de resolución es el proceso mediante el cual un identificador constituye una entrada (un pedido) a un servicio de red para obtener información acerca de un objeto digital. El sistema DOI posee un directorio central. Cuando un usuario hace clic sobre un DOI, un mensaje es enviado al directorio central donde una dirección web es asociada con dicho DOI. Esta ubicación se envía nuevamente al navegador del usuario con un mensaje especial que le dice al sistema que se dirija a "esa dirección particular de Internet". Cuando un objeto es movido a un nuevo servidor o cuando el propietario vende el producto a otra compañía, el cambio de URL es grabado en el directorio y todos los usuarios son direccionados al nuevo sitio web. De esta manera el cambio de URL de un documento sólo se hace en el directorio y no en todas las referencias hacia él (Martín, 2013).

2.6.8.2. Handle System

El Sistema *Handle* (o mejor conocido por *Handle System* en inglés), es un sistema de información distribuido diseñado para proporcionar un servicio de nombres global eficiente, extensible y protegido para su uso en redes como Internet. El *Handle System* incluye un protocolo abierto, un espacio de nombres y una implementación de referencia del protocolo. El protocolo permite que un sistema informático distribuido almacene nombres o maneje recursos digitales y resuelva dichos identificadores en la información necesaria para localizar, acceder y hacer uso de los recursos. Estos valores asociados se pueden cambiar según sea necesario para reflejar el estado actual del recurso identificado sin cambiar el identificador. Esto permite que el nombre del elemento persista sobre cambios de ubicación y otra información de estado actual. Cada identificador puede tener su propio administrador (o conjunto de administradores) y la administración se puede hacer en un entorno distribuido. El *Handle System* admite una resolución segura del *Handle* (entiéndase esto como los identificadores son asignados mediante este sistema, el cual es una referencia abstracta a un recurso). Los servicios de seguridad como la confidencialidad de los datos, la integridad de los datos y el no repudio son proporcionados con la solicitud del cliente. El *Handle System* proporciona un servicio de nombres confederado que permite que cualquier espacio de nombres local existente se una al espacio de nombres de identificadores global obteniendo una autoridad de nomenclatura de sistema *Handle* única. Los nombres locales y su (s) valor (es) se mantienen

intactos después de unirse al sistema Handle.

Cualquier solicitud de *Handle* al espacio de nombres local puede ser procesada por una interfaz de servicio que entienda el protocolo del sistema *Handle*. Combinado con la autoridad de nomenclatura única, cualquier nombre local se garantiza la unicidad bajo el espacio de nombres de identificador global.

Hay varios servicios utilizados hoy en día para proporcionar servicio de nombres para los recursos de Internet. Entre ellos, el Sistema de Nombres de Dominio (DNS) (Mockapetris, 1983, 1987) es el más utilizado. DNS está diseñado "para proporcionar un mecanismo para nombrar los recursos de tal manera que los nombres son asignables en direcciones IP y son utilizables en diferentes hosts, redes, familias de protocolos, Internet y organizaciones administrativas". El Sistema *Handle* ha sido diseñado desde el principio como como un servicio de nombres de propósito general. Está diseñado para dar cabida a un gran número de entidades y permitir la administración distribuida a través de la Internet pública. El modelo de datos del Sistema *Handle* permite que el control de acceso se defina al nivel de cada uno de los valores de datos asociados con un identificador dado. Más aún, cada identificador puede definir su propio conjunto de administradores que son independientes de la red o el administrador de host.

El sistema *Handle* está diseñado sobre los siguientes objetivos:

Unicidad : Cada *Handle* es único en todo el mundo dentro del Sistema Handle.

Persistencia : Los identificadores pueden utilizarse como identificadores persistentes para los recursos de Internet. Un identificador no tiene que derivarse de la entidad que nombra. Mientras que un nombre existente, o incluso un mnemónico, puede ser incluido en un identificador por conveniencia, la única conexión operacional entre un identificador y la entidad que nombra se mantiene dentro del sistema de control. Esto, por supuesto, no garantiza la persistencia, que es una función de la atención administrativa. Pero permite que el mismo nombre persista sobre los cambios de ubicación, propiedad y otras condiciones del estado. Por ejemplo, cuando un recurso con nombre se mueve de una ubicación a otra, el identificador puede mantenerse válido al actualizar su valor en el Sistema de control para reflejar la nueva ubicación.

Múltiples instancias : un único identificador puede hacer referencia a varias instancias de un recurso, en ubicaciones diferentes y posiblemente cambiantes en una red. Las aplicaciones pueden aprovechar esto para aumentar el rendimiento y la fiabilidad. Por ejemplo, un servicio de red puede definir múltiples puntos de entrada para su servicio con un solo identificador para distribuir la carga de servicio.

Múltiples Atributos : Un solo identificador puede referirse a múltiples atributos de un recurso, incluidos los servicios asociados, disponibles a través de cualquier método en ubicaciones de la red diferentes y posiblemente cambiantes. Así, los identificadores pueden utilizarse como puntos de entrada persistentes en un mundo en evolución de servicios asociados con recursos identificados.

Espacio de nombres extensible : Los espacios de nombres locales existentes pueden unirse al espacio de nombres del identificador mediante la adquisición de una autoridad de nomenclatura de identificador única. Esto permite que los espacios de nombres locales se introduzcan en un contexto global mientras se evita el conflicto con los espacios de nombres existentes. El uso de las autoridades de nomenclatura también permite la delegación de servicio, tanto de resolución como de administración, a un servicio local de manejo.

Soporte internacional El espacio de nombres de control se basa en *Unicode 3.0*, que incluye la mayoría de los caracteres utilizados actualmente en todo el mundo. Esto permite que los *Handles* se utilicen en cualquier entorno nativo. El protocolo del servicio obliga a utilizar UTF-8 como la codificación utilizada para identificadores.

Modelo de servicio distribuido : El sistema de control define un modelo de servicio jerárquico de tal manera que cualquier espacio de nombres de identificador local puede ser atendido por un servicio de identificador local correspondiente, por el servicio global o por ambos. El servicio global, conocido como G*lobal Handle Registry*, puede utilizarse para enviar cualquier solicitud de servicio de identificador al servicio local responsable. El modelo de servicio distribuido permite la replicación de cualquier servicio dado en varios sitios de servicio y cada sitio de servicio puede distribuir su servicio en un grupo de servidores individuales (tener en cuenta que local aquí se refiere sólo al espacio de nombres y a las preocupaciones

administrativas. Un servicio de control local de hecho podría tener muchos sitios de servicio distribuidos a través de Internet.)

Servicio de nombres protegidos : El Sistema *Handle* permite la resolución de nombres seguros y la administración a través de la Internet pública. El protocolo del Sistema *Handle* define los mecanismos estándares para la autenticación tanto del cliente como del servidor, así como la autorización del servicio. También ofrece opciones de seguridad para asegurar la integridad y confidencialidad de los datos.

Servicio de Administración Distribuida : Cada *Handle* puede definir su propio administrador/es o grupo de administrador/es. La propiedad de cada *Handle* se define en función de su administrador o grupo de administradores. Esto, combinado con el protocolo de autenticación de *Handle System*, permite que cualquier *Handle* sea administrado de forma segura a través de la red pública por su administrador en cualquier ubicación de red.

Servicio de resolución eficiente : El protocolo del sistema está diseñado para permitir un rendimiento de resolución de nombres altamente eficiente. Para evitar que la resolución se vea afectada por un servicio de administración costoso, las interfaces de servicio separadas (es decir, los procesos de servidor y sus puertos de comunicación asociados) para la resolución y administración de nombres pueden definirse por cualquier servicio *Handle*.

Probablemente sea mejor ver el Sistema *Handle* como un servicio de enlace de nombre-atributo con un protocolo específico para crear, actualizar, mantener y acceder de forma segura a una base de datos distribuida. Está diseñado para ser un servicio habilitado para compartir información segura y recursos a través de redes como Internet. Las aplicaciones del Sistema *Handle* podrían incluir servicios de metadatos para publicaciones digitales, servicios de administración de identidad para identidades virtuales o cualquier otra aplicación que requiera resolución y/o administración de identificadores globales únicos.

Cada *Handle* consta de dos partes: la autoridad de denominación, también conocida como su prefijo, y un nombre local único bajo la autoridad de nomenclatura, también conocido como su sufijo:

```
<Handle> ::= <Handle Naming Authority> "/" <Handle Local Name>
```

La autoridad de nomenclatura y el nombre local están separados por el carácter ASCII "/". La colección de nombres locales bajo una autoridad de nombres define el espacio de nombres de identificador local para esa autoridad de nomenclatura. Cualquier nombre local debe ser único en su espacio de nombres local.

La singularidad de una autoridad de nomenclatura y un nombre local bajo esa autoridad asegura que cualquier *Handle* sea globalmente único dentro del contexto del Sistema *Handle*.

Por ejemplo, "10.1045/january99-bearman" es un *handle* para un artículo publicado en la revista *D-Lib*. Su autoridad de nomenclatura es "10.1045" y su nombre local es "january99-bearman". El espacio de nombres *handle* puede considerarse un superconjunto de muchos espacios de nombres locales, y cada espacio de nombres local tiene una autoridad de nomenclatura única bajo el Sistema *Handle*.

La autoridad de nomenclatura identifica la unidad administrativa de creación, aunque no necesariamente la administración continua, de los identificadores asociados. Se garantiza que cada autoridad de nomenclatura es globalmente única dentro del sistema de control. Cualquier espacio de nombres local existente puede unirse al espacio de nombres de identificador global obteniendo una autoridad de nomenclatura única para que cualquier nombre local bajo el espacio de nombres se pueda referenciar globalmente como una combinación de la autoridad de nombres y el nombre local como se muestra arriba (Sun et al., 2003).

La Fundación Internacional DOI (IDF, por sus siglas en inglés) apoya firmemente el *Handle System* y cree que es el mejor componente de infraestructura actualmente disponible para administrar objetos digitales. Es por eso que el sistema DOI utiliza el *Handle System*. Los *Handles* por sí mismos son necesarios, pero no suficientes para la función del sistema DOI, un marco completo para administrar el contenido intelectual y facilitar el comercio electrónico. Si bien el sistema DOI se basa en el Sistema *Handle*, uno podría preguntarse porque no utilizar simplemente *Handles*, pero esto no es del todo acertado pues los nombres DOI son más que *Handles* (Factsheet, 2015).

Los "Handles" (entiéndase esto como los identificadores asignados mediante este sistema, el cual es una referencia abstracta a un recurso), ya

sea directamente o como parte del Sistema DOI[50], se utilizan actualmente para identificar muchos artículos en Internet incluyendo publicaciones, artículos, autores y cantidades crecientes de datos de investigación a través de esfuerzos e iniciativas como es el caso de *Datacite*[51].

Los DOIs (que utilizan el sistema *Handle* para la resolución de identificadores) se han normalizado bajo la norma "*ISO 26324 Information and documentation - Digital object identifier system*" (Reilly, 2013).

El *Handle System* proporciona un servicio de nombres global de propósito general que permite la resolución segura de nombres a través de Internet, diseñado para permitir a un amplio conjunto de comunidades utilizar dicha tecnología para identificar contenido digital independientemente de la ubicación. El sistema DOI utiliza el Sistema *Handle* como un componente en la construcción de una aplicación de valor agregado, para la identificación persistente, semánticamente interoperable, de las entidades de propiedad intelectual.

Como se mencionó previamente, en la actualidad existen otras tecnologías para la resolución de nombres, tal es el caso de DNS o URN (Uniform Resource Name). Si bien el DNS resuelve el nombre de un dominio asignando una dirección IP única para ello, para localizar un sitio web, sirve sólo y únicamente para este propósito (DNS es un excelente mecanismo de resolución de nombres de dominio, pero esto no lo convierte en un mecanismo de resolución de ningún tipo para otros nombres o identificadores), sin embargo presenta ciertas limitaciones en relación a la gestión de identificadores. En este sentido, el sistema *Handle* se diseñó como un sistema de resolución de objetos digitales y sirve como un nivel de indirección, a cualquier tipo de datos del estado actual, que se desee asociar con el objeto a través del mecanismo de resolución de identificadores.

El *Handle System* proporciona una forma de utilizar DNS y URLs para identificadores, que a la vez proporciona un identificador que se puede resolver sin utilizar DNS y URLs, si decide utilizarlo de esa manera(Factsheet, 2015). Este enfoque utilizará el actual sistema de resolución *Handle* y/o DOI, estándares e infraestructura para permitir la resolución de identificadores basados en *Handles* o DOIs utilizando el Protocolo *Handle* o mediante servicios web encargados de resolver los

[50] http://www.doi.org
[51] http://datacite.org

mismos identificadores. Ejemplos de servicios que pueden resolver *Handles* y DOIs son `http://hdl.handle.net` y `http://dx.doi.org`.

Resolver un *Handle* o DOI puede producir una URL que se puede seguir para obtener más información, o podría proporcionar información acerca del objeto que se identifica directamente sin necesidad de utilizar HTTP en absoluto. Los identificadores basados en el sistema *Handle* se pueden resolver prácticamente para cualquier cosa, incluyendo descripciones de texto (en cualquiera o en todos los idiomas) o referencias adicionales. Los identificadores se pueden resolver directamente a los datos a los que se refieren, prácticamente en cualquier formato, o pueden resolverse a otras referencias, como URL HTTP, direcciones de correo electrónico, claves públicas u otros datos estructurados. El sistema de resolución ha sido diseñado para la seguridad, la escalabilidad y la eficiencia, mientras que soporta un espacio de nombres relativamente plano de identificadores el cual es esencial para permitir la persistencia(Reilly, 2013).

Otra iniciativa basada en la capacidad y funcionalidad del *Handle System* es *ePIC* fundada en 2009 por un consorcio de socios europeos con el fin de proporcionar servicios PID (Persistent Identifiers) para la *Comunidad Europea de Investigación* (European Research Community) (ePIC, 2017)

2.6.8.3. ORCID

Open Researcher and ContributorID (ORCID) es una organización internacional, interdisciplinaria, abierta y sin fines de lucro creada para resolver el problema de la identificación, ambigüedad y duplicidad en los nombres de los investigadores (autores y colaboradores) mediante la creación de un registro único de identificadores persistentes, en beneficio de todas las partes interesadas, incluidas las instituciones de investigación, organizaciones financiadoras, editores, y los propios investigadores. A diferencia de DOI, no existe una agencia de registro ni es necesario pagar para crear un registro, ORCID es independiente, con diferentes *stakeholders* o partes interesadas, no limitada a un área geográfica ni temática, transparente y basada en código abierto, sin ninguna dependencia editorial, de aplicación global, y permite enlazar los registros de sus usuarios con otros sistemas de identificación de autores como *Scopus Author ID* y *ResearcherID* (Thomas et al., 2015; García-Gómez, 2012; Haak et al.,

2012).

En contraposición a esta alternativa, existen otras que proveen la mayor parte de las veces un alcance local y no rebasa los límites de una base de datos, una organización o un país, tal es el caso del *Scopus Author ID*, de *Elsevier* (Andalia et al., 2015). *Scopus Author ID* es otro identificador utilizado específicamente por la base de datos *Scopus* y posee muchas de las funcionalidades que ofrece *ResearcherID*, en el sentido que este identificador ayuda a administrar la lista de citas y publicaciones. No es necesario registrarse para obtener un *Scopus Author ID*, si un autor posee un artículo indexado en la base de datos, la plataforma asigna automáticamente un *Scopus Author ID* (The University of Chicago Library , 2016), es necesario aclarar que el ámbito de aplicación de este identificador se circunscribe únicamente a esta base de datos, es decir, no es posible asociar publicaciones de otras fuentes que no sean las indexadas por *Scopus*. Esta base de datos también permite unificar las diversas formas con las que aparece el nombre de un mismo autor, así mismo permite vincular los datos de autor y la lista de publicaciones de *Scopus* con ORCID, y viceversa (The University of the Sunshine Coast Library, 2016), en consecuencia, Scopus distingue entre autores con el mismo nombre asignando a cada autor un identificador distinto y agrupando todos los documentos escritos por ese autor (University of Tasmania, 2016).

ORCID (ORCID, 2015) vio la luz en 2012, impulsada por algunos de los editores comerciales académicos, bibliotecas nacionales, sociedades profesionales y principales repositorios de acceso abierto más importantes, cuyo objetivo es crear un registro centralizado de todos los "investigadores y colaboradores" de productos académicos, lo que permite identificadores únicos que eliminan la ambigüedad respecto a la identificación de sus contribuciones. ORCID en primer lugar será capaz de identificar a todos los autores y colaboradores por su documento; y en segundo a las instituciones y los individuos que dispongan de perfiles con los datos recogidos por ORCID sobre ellos, pudiendo sincronizarlos y actualizarlos con sus perfiles personales en ORCID y otros perfiles que puedan tener en otros lugares (Arévalo and Vázquez, 2016).

Este genial identificador está compuesto por 16 dígitos, construido sobre la base de la norma ISO 27729:2012, que permite a los investigadores disponer de un código de autor permanente e inequívoco que distingue con precisión tanto su producción como su quehacer científí-

fico. Los investigadores pueden registrarse individualmente en ORCID de forma gratuita con el propósito tanto de obtener su código como de almacenar, documentar y gestionar su labor profesional. También facilita la colaboración a partir de la identificación de otros especialistas con intereses similares. Una vez obtenido el registro ORCID, es posible añadir información relativa al autor como: correo electrónico, la identificación normalizada (uniforme) tanto del autor como de la institución donde trabaja, posibles variantes del nombre estandarizado, así como las referencias de sus publicaciones, entre otros datos (Andalia et al., 2015).

ORCID aborda un problema real: la vinculación de forma fiable de los autores con los resultados de la investigación y ayudando a estos a recibir crédito por su trabajo. El identificador único también permite a las instituciones automatizar el intercambio de información con otras organizaciones incrementando de este modo la calidad de los datos. Esto ahorra tiempo a los académicos y dinero a las instituciones, además resulta de utilidad al momento de intentar obtener promociones, pues compartiendo el ORCID (y la lista de publicaciones vinculadas a este identificador) la institución ya puede contar con el historial completo del interesado.

En la presentación de trabajos, los autores comparten su ORCID y otra información relevante con el editor. Tras la aceptación, los editores integran esta información en los metadatos de salida. Entonces los metadatos se desplazan automáticamente a través de los sistemas relevantes: editorial, ORCID, institución de investigación y fuentes de financiación. ORCID no se limita a artículos de revistas. Los principios se pueden aplicar a cualquier tipo de resultados y trabajos (conjuntos de datos, código fuente, tesis, patentes) e incluso como antecedentes, tales como historial de empleo (Reimer, 2015).

ORCID trabaja con más de 400 organizaciones de investigación, editores, proveedores de fondos y asociaciones profesionales para incrustar el identificador ORCID en varios flujos de trabajo de investigación, tales como la presentación de las propuestas de subvención a los organismos de financiación, los manuscritos a los editores de revistas y los conjuntos de datos a los repositorios de datos (Akers et al., 2016).

La idea de contar con este identificador único es que, por un lado sea de aplicación universal permitiendo no sólo identificar a un autor sin importar el lugar de trabajo, el idioma de publicación, la revista donde

publique, el equipo de colaboradores, sino también poder gestionar de forma única el conjunto completo de publicaciones (materiales en blog personales y sitios Web, artículos y libros, programas docentes, software, entre otros) de su autoría o coautoría. Y, por otro lado este registro único permitirá la automatización de los procesos encargados de contabilizar la productividad de un autor o científico, entregando como resultado un indicador mucho más preciso sin la asignación errónea de créditos que no corresponden (sumando registros que no son de la autoría del investigador en cuestión) ni "olvidando" trabajos (algunos trabajos que si son de la autoría del involucrado, pero por problemas como la imposibilidad de saber si un trabajo es del autor que se está evaluando, no son contabilizados).

Sin embargo la sola aplicación de este mecanismo no es suficiente, pues como se mencionó con anterioridad, la sola identificación del autor no resuelve todo el espectro del problema, la indexación de publicaciones duplicadas escapa del alcance de ORCID y es cuando iniciativas como DOI cobran aún más relevancia. Y en este mismo sentido la utilización de identificadores como el ISBN (International Standard Book Number)[52] para los libros y el ISSN (International Standard Serial Number)[53] para revistas podrían brindar aún mejores resultados.

ISNI (International Standard Name Identifier)[54] es el número estándar global certificado por ISO (ISO 27729) para identificar a los millones de contribuyentes de obras creativas y a los que participan en su distribución, incluidos investigadores, inventores, escritores, artistas, diseñadores gráficos, intérpretes, productores, editores, intermediarios y más. Forma parte de una familia de identificadores de normas internacionales que incluye identificadores de obras, grabaciones, productos y titulares de derechos en todos los repertorios, por ejemplo: *DOI, ISAN, ISBN, ISRC, ISSN, ISTC* y *ISWC*.

La misión de la Autoridad Internacional ISNI (ISNI-IA) es asignar al nombre público de un investigador, inventor, escritor, artista, intérprete, editor, etc., un número único de identificación persistente para resolver el problema de ambigüedad del nombre tanto en la búsqueda como en la recopilación de información; y difundir cada ISNI asignado a través de todos los repertorios de la cadena de suministro global para que

[52] http://www.isbn.org/
[53] http://www.issn.org/
[54] http://www.isni.org/

cada trabajo publicado pueda atribuirse inequívocamente a su creador dondequiera que se describa ese trabajo.

Al lograr estos objetivos, el ISNI actuará como un identificador de puente entre múltiples dominios y se convertirá en un componente crítico en aplicaciones de enlaces de datos y aplicaciones basadas en la Web semántica.

Por lo mencionado anteriormente, ISNI parece tener "casi" el mismo propósito que ORCID, sin embargo ORCID e ISNI son organizaciones separadas que se ocupan de diferentes aspectos de la identificación inequívoca de personas y partes. El contexto, los antecedentes y los objetivos de cada organización son distintos.

ORCID se estableció para resolver el problema de la atribución exacta de los resultados de la investigación académica a los investigadores individuales. El sistema ORCID se basa en la colaboración entre editoriales, universidades, organismos de financiamiento, investigadores y otras partes interesadas en las comunicaciones académicas. ORCID se compromete a permitir a los investigadores individuales crear, reclamar, gestionar y controlar la privacidad de sus datos u, opcionalmente, delegar la gestión de sus datos a su universidad u otro tercero.

Debido a que ORCID y ISNI tienen diferentes propósitos y sirven a diferentes comunidades, ambas organizaciones son necesarias. Las organizaciones tendrán diferentes datos, tendrán diferentes reglas de privacidad y propiedad para los datos, tendrán diferentes modelos de negocio y ofrecerán diferentes servicios. Lo más importante de todo es que ISNI y ORCID estarán identificando diferentes cosas para diferentes comunidades.

ORCID está comprometida a ser interoperable con otros esquemas de identificadores, incluyendo ISNI. Con este fin, ORCID e ISNI están coordinando sus esfuerzos cuando se superponen en las comunidades académica y de investigación. Los identificadores ORCID utilizan un formato compatible con la norma ISNI ISO, por ello ISNI ha reservado un bloque de identificadores para su uso por ORCID, por lo que no habrá superposiciones en las asignaciones de identificadores. Ambas organizaciones están trabajando juntas para considerar oportunidades adicionales de colaboración (Clark, 2012).

2.6.8.4. GRID

GRID (*Global Research Identifier Database*)[55] es un conjunto de datos, abiertamente accesible, de información sobre organizaciones de investigación a nivel mundial. Lanzada en 2015, es una base de datos en línea, sin cuotas anuales, sin complejos acuerdos de licencia, totalmente gratuita y fácil de usar que abre información sobre organizaciones de investigación de todo el mundo a científicos encargados de recopilar y estudiar datos, desarrolladores e innovadores dentro de organizaciones académicas y comerciales. El conjunto de datos GRID aborda el problema de los datos desordenados e incoherentes sobre las instituciones de investigación, asegurando que cada entidad se incluye correctamente y sólo una vez.

La fuente de datos más frecuente de referencias institucionales son los datos de afiliación de autores que se encuentran en artículos científicos. Típicamente, los creadores de artículos originales de investigación y de actas de congresos, incluyen en sus artículos los datos del departamento, organización y dirección a la que pertenecen. Esto no sólo permite a los lectores reconocer el origen de la investigación, sino que también sirve como un mecanismo para agregar y evaluar la producción científica con el fin de proporcionar métricas de la ciencia.

Sin embargo, al adquirir y procesar grandes cantidades de afiliaciones de autores, se hace evidente que una variación significativa en el formato y la estructura evita la agregación correcta y la presentación de informes efectivos. Esto, junto con los cambios en el nombre y la estructura institucional a través del tiempo, hace que la integración a gran escala de estos datos sea prohibitivamente costosa, dado el esfuerzo manual requerido para desambiguar apropiadamente cada afiliación.

GRID proporciona un servicio de desambiguación automática para superar estos desafíos. Al explotar la riqueza de datos que se han adquirido durante el procesamiento de los datos de adjudicación y publicación, junto con una extensa base de datos de instituciones, se puede proporcionar coincidencia algorítmica de cadenas de afiliación de autor a las instituciones (GRID, 2017a).

La base de datos en línea contiene 50.000 nombres institucionales manualmente seleccionados, junto con identificadores únicos e informa-

[55] https://www.grid.ac

ción de geolocalización en 212 países. Los datos se derivan de fuentes de financiación y fuentes de publicación abiertamente accesibles, como *NIH reporter*, *PubMed* y la *UK Gateway to Research* (Wheeler, 2015).

GRID ha sido ampliamente adoptado en las empresas de la cartera de *Digital Science*[56] para facilitar el intercambio de datos, aumentar la funcionalidad y apoyar nuevas características. Bajo la creencia que estos beneficios deberían compartirse más ampliamente en la comunidad científica para fomentar la innovación y aumentar la interoperabilidad. Empresas como *Altmetric*, *Figshare*, *Elements* y *Dimensions* utilizan la plataforma actualmente (GRID, 2017b).

Los registros GRID contienen una gran cantidad de metadatos obtenidos de fuentes confiables. Entre éstas se incluyen fechas establecidas, alias de nombre, siglas y geolocalización. Además se incluyen enlaces a páginas web externas como *Wikipedia* y sitios web oficiales, así como identificadores externos como *ISNI*, *GeoNames*[57] y *Fundref*[58] (Meddings, 2017) (observar Figura 2.18). En esta base se pueden encontrar los siguientes tipos de instituciones: Educativas, Cuidado de la Salud, Compañías, Archivos, Organizaciones sin fines de lucro, Gubernamentales, Instalaciones dedicadas a la investigación y otros (GRID, 2017).

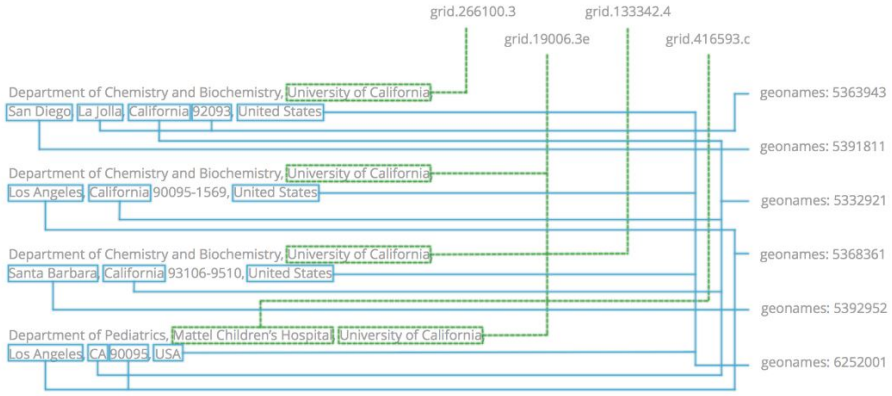


Figura 2.18: Esquema de los identificadores propuestos por GRID

[56] https://www.digital-science.com/
[57] http://www.geonames.org/
[58] http://search.crossref.org/funding

2.7. Visualización de Información

En las décadas anteriores era muy común que los datos se presentaran textualmente o mediante gráficos estáticos, en estos casos la información representada estaba limitada a cantidades pequeñas, pero en los últimos años estos tipos de representaciones se han tornado poco útiles cuando se trata de conjuntos de datos que contienen millones de elementos de datos (Keim, 2002). Los sistemas actuales almacenan grandes cantidades de datos y al no tener la posibilidad de explorarlos adecuadamente, los datos se vuelven inútiles y las bases de datos se convierten en meros "depósitos".

La reciente aparición de interfaces gráficas ha permitido una interacción directa con la información visualizada, dando lugar a más de una década de investigación en visualización de información (*InfoVis*) (Heer et al., 2005). *InfoVis* busca aumentar el conocimiento humano mediante el aprovechamiento de las capacidades visuales humanas para dar sentido a la información abstracta (Card et al., 1999), proporcionando los medios por los cuales los seres humanos mediante sus capacidades perceptivas, pueden lidiar con el constante aumento de la cantidad de datos disponibles. Una definición un tanto similar es la aportada por (John and Joseph, 1998), que define la *Visualización de Información* como *"la práctica de mapear conjuntos de datos dentro de medios de comunicación visual con el fin de ayudar a los usuarios en la exploración de estos conjuntos de datos o la comunicación al respecto a los demás"*.

El objetivo de *InfoVis* es profundizar en los datos o conceptos ocultos. A menudo la información se oculta simplemente por la enorme cantidad de datos disponibles. De este modo, la *Visualización de Información* también puede ser vista como un convertidor entre los datos subyacentes y la percepción humana de la misma (Prinz, 2006). El representar grandes cantidades de información mediante abstracciones no es una tarea fácil ya que el usuario no tiene ninguna idea preconcebida de cómo estos datos pueden ser representados. Así, los métodos de *InfoVis* deben ser capaces de hacer frente a los datos que parecen ser al azar, pero aún así contiene información valiosa (Kosara et al., 2003).

Las representaciones gráficas de *Visualización de Información* no son un invento de la era de la computación y de los gráficos por ordenador. Los ejemplos históricos de *Visualización de Información* se remontan a las pinturas rupestres (Ware, 2004). En dichas pinturas se visualiza

el proceso de la caza o cualquier otra rutina diaria como se observa en la Figura 2.19 (imagen tomada del libro *Information Visualization - Perception for Design 2nd ed.* (Ware, 2004)).

Figura 2.19: Pintura rupestre

El argumento estándar para la visualización es que la explotación del procesamiento visual puede ayudar a explorar o explicar los datos. El campo de la visualización es un campo activo en constante evolución, debido a que los desafíos de diseño son importantes y muchas veces no se llegan a comprender en su totalidad (Munzner, 2000).

La idea básica de la exploración visual de los datos es la de presentar los datos en alguna forma visual, permitiendo que los humanos puedan obtener conocimiento, sacar conclusiones, e interactuar directamente con los mismos. Con este tipo de representaciones basadas en grandes cantidades de datos, los usuarios pueden *detectar patrones o comportamientos que se deseaban evaluar, como así también descubrir comportamientos y relaciones entre los datos desconocidos hasta el momento.* Además de la participación directa del usuario, las principales ventajas de la exploración visual de los datos según (Keim, 2002), son:

- La exploración de datos visuales pueden tratar con datos muy heterogéneos y con ruido.

- La exploración visual de los datos es intuitiva y no requiere la comprensión de complejos algoritmos matemáticos o estadísticos.

2.7.1. Procesamiento preatentivo

Los seres humanos tienen notables capacidades de percepción (Andrews, 2011b):

- Analizar, reconocer y recordar las imágenes rápidamente.
- De forma rápida y automáticamente detectar patrones y cambios en el tamaño, color, forma, movimiento, o la textura.

Interfaces basadas en texto requieren esfuerzo cognitivo para entender su contenido informativo. La *Visualización de Información* tiene por objeto presentar la información visualmente, en esencia, para reducir la carga de trabajo cognitivo al sistema perceptivo visual humano.

Un principio cognitivo fundamental a tener en cuenta cuando se realiza el procesamiento de información, es si dicho procesamiento se hace de forma deliberada o pre-consciente (Munzner, 2000). Parte de la información visual de bajo nivel se procesa de forma automática por el sistema perceptivo humano sin el foco de atención consciente. Este tipo de tratamiento se llama automático, preatentivo, o selectivo. Un ejemplo del tratamiento preatentivo es el indicado en (Ware, 2004), en el cual se intenta identificar unas cerezas en medio de unas hojas como se observa en la Figura 2.20 (imagen tomada de (Ware, 2004)), este ejemplo además muestra la importancia de contar con la visión del color a la hora de realizar un análisis visual. En la imagen de la izquierda es difícil de encontrar las cerezas, en cambio en la figura de la derecha se produce el efecto visual *pop out*, y las cerezas saltan a la vista.

La explotación del procesamiento precognitivo es deseable en un sistema de visualización para que los recursos cognitivos puedan ser liberados para otras tareas. Muchas de las funciones pueden ser preatentivamente procesadas, incluida la duración, la orientación, el contraste, la curvatura, la forma y el tono (Triesman and Gormican, 1988). Sin embargo, el procesamiento preatentivo posee la limitación de trabajar con una sola característica en un caso dado, por lo que la mayoría de las búsquedas realizadas con un conjunto de más de una característica no son precognitivo. Por ejemplo, los cuadrados verdes entre los cuadrados y círculos

de color rojo y verde no saldrán a la vista, y pueden ser descubiertos por un proceso de búsqueda consciente mucho más lento, como se ve en la Figura 2.21 (imagen tomada de (Ware, 2008)).

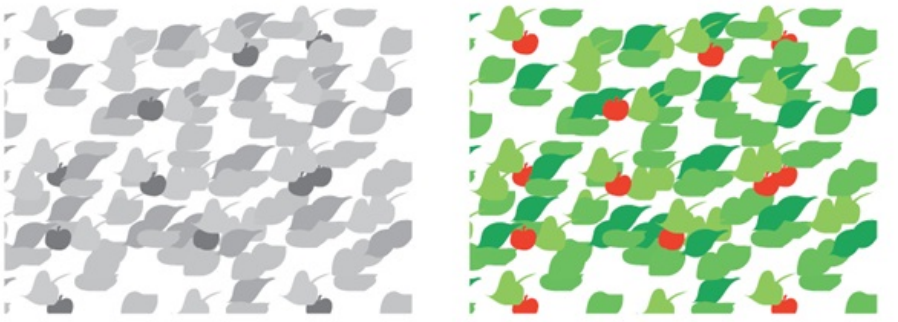


Figura 2.20: Procesamiento con preatención y utilizando la visión del color

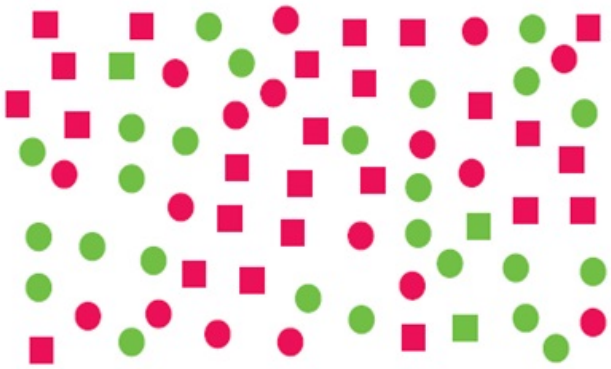


Figura 2.21: Procesamiento sin preatención

La comprensión de las bases de la transformación de los datos en "entendimiento", conocido como el proceso de *InfoVis*, es crucial para desarrollar estrategias eficaces que ayuden a los usuarios a alcanzar sus metas informativas. Se han presentado varias aproximaciones conceptuales a este proceso, todos ellos convergen en la definición de tres pasos principales (Pascual Cid, 2010):

- *De los Datos a la Información*: Una vez que se han recolectado los datos pertinentes sobre un problema objetivo, debe estructurarse y organizarse para transformarse en información. Se pueden identificar tres tareas relacionadas con dicha conversión:

  - la *recogida* y *almacenamiento* de datos brutos pertinentes al objeto de estudio
  - el *tratamiento* y *transformación* de dichos datos para filtrar errores. Este proceso implica la supresión de registros irrelevantes y redundantes, y la creación de magnitudes derivadas.
  - el *uso de metadatos* para construir tablas de datos organizadas según su significado.

- *De la Información a la Representación Visual*: Con los datos ya transformados en información, este paso trata de la conversión de la información en una representación perceptiva, principalmente en forma visual. Esta representación perceptiva tiene que mostrar la información de tal manera que los patrones y estructuras subyacentes tienen que ser fácilmente identificados. Se debe aplicar un esquema de notación, que es un lenguaje visual particular que mapea información en gráficos, para aprovechar los conocimientos previos o la experiencia del usuario. Hay dos clases de representaciones, la arbitraria convencional y la sensorial. La primera es la aprendida en el tiempo por el receptor y su característica principal es que no tiene ninguna base perceptiva. La última se basa en símbolos y aspectos de visualización que utilizan la capacidad perceptiva del cerebro sin ningún conocimiento o experiencia previa. Para que la visualización produzca un impacto perceptual para el usuario, el diseñador debe tener en cuenta tanto las representaciones convencionales sensoriales como arbitrarias. Además, la percepción visual, la psicología cognitiva e incluso la lingüística deben considerarse para proporcionar una experiencia visual agradable y comprensible.

- *De la Representación Visual a la Comprensión*: Una vez que la representación visual se ha construido, tiene que ser presentada al receptor. Con el fin de ayudar a obtener entendimiento y construir conocimiento, la visualización debe permitir al usuario interactuar con él y potenciar el discurso analítico.

2.7.2. Visualización interactiva

InfoVis abarca las técnicas de visualización que tienen que ver principalmente con datos abstractos, es decir, los datos para los cuales el usuario no tiene un modelo mental preconcebido. Por esta razón, la interacción es especialmente importante en *InfoVis*, ya sea para la exploración, análisis y/o presentación de los datos (Kosara et al., 2003). La interacción permite al usuario implícitamente formar modelos mentales de las correlaciones y las relaciones entre los datos, a través del reconocimiento de patrones, marcando y centrándose en esos patrones, la formación de hipótesis y pruebas mentales.

La utilización de representaciones estáticas como imágenes o gráficos sin interacción han quedado atrás, el usuario de hoy en día necesita interactuar con la representación presentada, la herramienta de soporte debe brindar las facilidades para intercambiar formas, colores y modos de representar la misma información. El campo de la informática que se ocupa del análisis de la interacción entre los seres humanos y el ordenador se llama Interacción Persona Ordenador (o HCI por sus siglas en ingles, *Human-Computer Interaction*) (Andrews, 2011a). Dado que el sentido primario del ser humano es el sentido de la vista, la mayor parte de la información puede ser transportada utilizando este canal.

Algunos tipos de interacción intentan imitar el mundo real o algunos aspectos de la naturaleza, ya sea con representaciones sencillas en 2D o con representaciones 3D un tanto más complejas, donde el usuario puede manipular los objetos representados. El uso de los ordenadores permite ir un paso más allá de simples representaciones del mundo real, permitiendo hacer agrupaciones o asociaciones impensables, o distorsiones sobre dichas representaciones proporcionando un mayor nivel de abstracción, aspecto fundamental de las técnicas de *InfoVis*.

Por último destacar que *InfoVis* es una área nueva y en constante avance, muchos de los expertos y autores del tema aun no se ponen de acuerdo en cuanto a algunas definiciones, sin embargo, uno de los aspectos a tener en cuenta y muy aceptado dentro de la comunidad es lo que Ben Shneiderman denomina el "mantra" de la visualización: *"Overview first, zoom and filter, then details on demand"* (Shneiderman, 1998). Según el autor, en estas cuatro reglas se sintetizan los aspectos a cubrir en todo proceso de búsqueda de información y en las interfaces que los soportan. Este mantra sugiere la necesidad de proporcionar una visión

general de los datos, que pueden ser filtrados, ampliados y modificados en cualquier momento para que el usuario pueda obtener una visión más profunda.

2.7.3. Visualizaciones actuales

La Visualización de Información cumple un papel relevante al realizar un análisis basado en citas bibliográficas. Poder plasmar mediante una representación, la información recolectada se vuelve una tarea compleja no solo según aumente la cantidad de información, sino también según se incremente el número de dimensiones objeto de estudio. En la actualidad existen un pequeño número de visualizaciones relacionadas con esta temática, son los motores de búsquedas, bases de datos bibliográficas o herramientas de análisis las encargadas de ofrecerlas.

Scopus presenta un conjunto de visualizaciones que permiten ver desde distintos puntos de vista los resultados obtenidos a partir de una búsqueda por autor (solo se indican las representaciones para este tipo de búsqueda ya que es de interés para este trabajo). Para un búsqueda dada (se toma como ejemplo una búsqueda por autor con el nombre *Torres-Salinas Daniel*), una vez realizada esta búsqueda (observar Figura 2.22), en la parte superior existe una opción llamada "Analyze author output", esta abre una nueva sección donde se puede apreciar las siguientes representaciones: "Documents" que presenta cuatro opciones *by source, by type, by year y by subject area* como se observa en la Figura 2.23, en dicha opción se agrupan los documentos publicados por origen (revista, congreso,...), por tipo de documento (artículo, nota, conferencia,...), por año de publicación y por área temática; la otra representación es "h-index" (observar Figura 2.25); la siguiente representación es "Citations" (Figura 2.24) la cual es un gráfico de barras que indica con diferentes colores la cantidad de citas recibidas por cada año y por último "Co-authors" que presenta un listado de todos los coautores para el autor objeto de la búsqueda.

La única visualización relacionada con el número de citas y con indicadores bibliométricos, es el *h-graph*(Figura 2.25). Este gráfico mide el impacto de un conjunto de artículos y muestra el número de citas por documento. Para poder acceder a este gráfico, dentro del panel de resultados, se debe elegir la opción "View h-graph" (aquí se muestra por medio de un par de curvas la intersección entre la cantidad de documentos y el

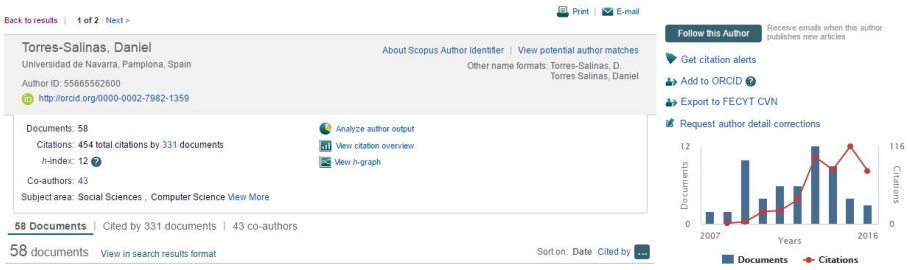


Figura 2.22: Resultados de la consulta por Autor en *Scopus* para Daniel Torres-Salinas

número de citas recibido, para poder calcular el *h-index*)

Los gráficos aquí presentados no dejan de ser sencillos, aunque ofrezcan cierto grado de interacción con el usuario, no dejan de estar muy separados unos de otros, es decir, cada gráfico muestra por separado una dimensión y cuentan solo una parte de la historia.

Por su parte la *WoS* presenta visualizaciones para analizar los resultados entregados, la más importante de ellas es sin dudas el "Citation Map" es una representación gráfica que muestra las relaciones de las citas (tanto las referencias citadas como los artículos que citan) entre una publicación y otras publicaciones utilizando diversas herramientas y técnicas de visualización. Con esta herramienta se puede analizar qué investigadores citan ciertos trabajos. También se puede optar por organizar y codificar mediante colores los resultados por autor, año, revista, área o tema, entre otras opciones. De la misma forma se puede establecer una representación gráfica de los trabajos que se han citado en una publicación. El *Citation Map* permite entre otras cosas visualizar las conexiones entre las citas y los trabajos publicados, y descubrir relaciones mucho más amplias. No se debe olvidar que esta representación es para una publicación individual, no se pueden visualizar un conjunto de publicaciones.

Sin embargo, al momento de escribir este trabajo, dicha representación no es visible con un navegador actualizado pues la *WoS* impone ciertas restricciones (como se puede ver en la Figura 2.26) limitando la utilización a navegadores totalmente desactualizados, por ejemplo, el navegador utilizado para realizar esta investigación es *Chrome* Versión

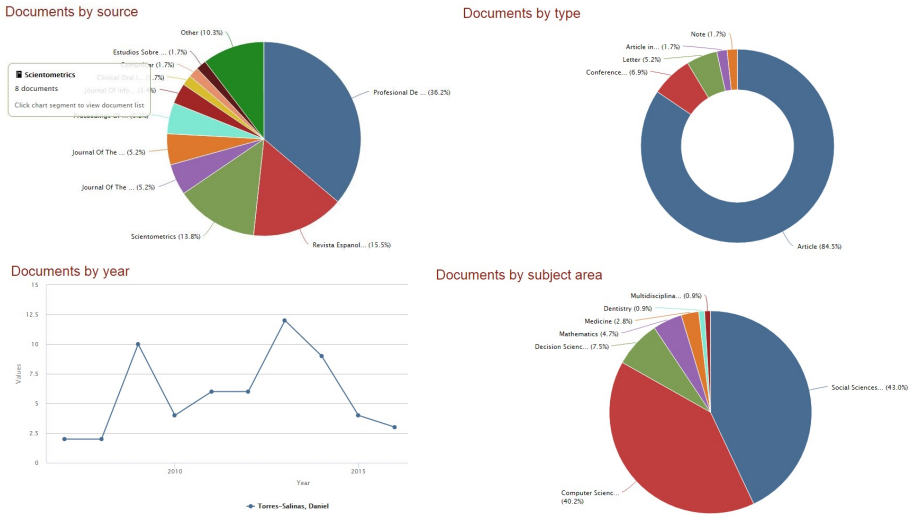


Figura 2.23: Las cuatro opciones de visualización de Documentos en *Scopus*

54.0.2840.99 m, siendo que la versión probada y recomendada por la *WoS* es la 43.

Lo mismo sucede con el buscador académico *Microsoft Academic Search*, quien en su primer versión ofrecía distintas visualizaciones, a saber: *Co-author Graph*, *Co-author Path* o *Citation Graph*.

El *Co-author Graph* (grafo de coautor) es una red/grafo que representa el top 30 de los coautores que han colaborado con el autor buscado. Este grafo permite observar las distintas interacciones entre los distintos autores, en los arcos que unen los nodos se puede observar la cantidad de trabajos en conjunto entre dos autores, y además permite acceder a dichas publicaciones.

El *Co-author Path* muestra como dos investigadores dados están conectados entre sí en lugar de centrarse en un único investigador, esta visualización ayuda a descubrir el camino, ruta o dirección que relaciona a estos dos investigadores analizando sus colaboradores.

Por último el *Citation Graph* es una herramienta similar al *Co-author Graph*, solo que muestra los autores que han citado al autor objeto de estudio. En el arco que une los dos nodos se observa la cantidad de veces que fue citado dicho autor y al hacer clic sobre este número se puede

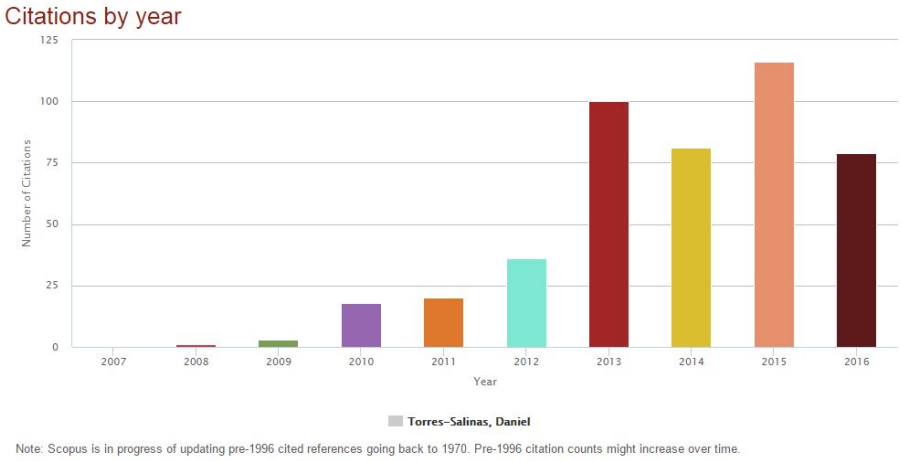


Figura 2.24: Gráfico de barras de las citas recibidas por año para un autor en *Scopus*

acceder a dichas citas.

Como se indicara en un principio, estas visualizaciones solo permanecen en la versión desactualizada y casi obsoleta del motor académico (MAS), en la nueva versión (MA) ya no se encuentran disponibles, puesto que estaban basadas en el complemento *SilverLight* de *Microsoft* el cual no es soportado por todos los navegadores.

Por último *Scholarometer*, ofrece un conjunto de visualizaciones para explorar los resultados obtenidos. La primera de ellas es el "Author Network" (o red/grafo de autor). Es una visualización en forma de grafo similar a *Co-author Graph* de MAS, en esta se ve la red de colaboración de los autores relacionados al autor objeto de la búsqueda como se observa en la Figura 2.27. Al posar el mouse sobre un nodo se ve el detalle de los indicadores y de los campos del conocimiento involucrados .

La siguiente visualización es *Discipline Network*, la cual es un grafo donde se encuentran todas las disciplinas posibles de los campos indicados en el criterio de búsqueda. Se representan las distintas relaciones entre una y otra, y además se pueden filtrar por el nombre de algún campo, al hacer esto en el grafo aparecen los nodos resaltados que coinciden con el nombre o parte del nombre del criterio ingresado como se observa en la Figura 2.28.

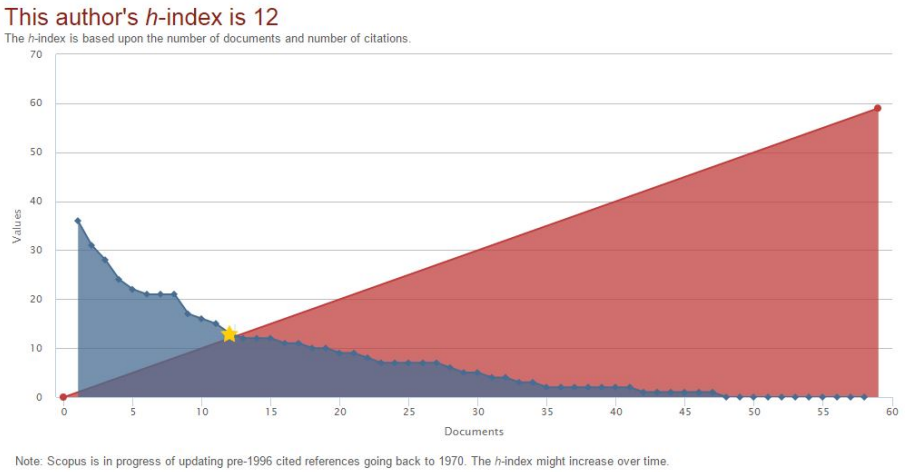
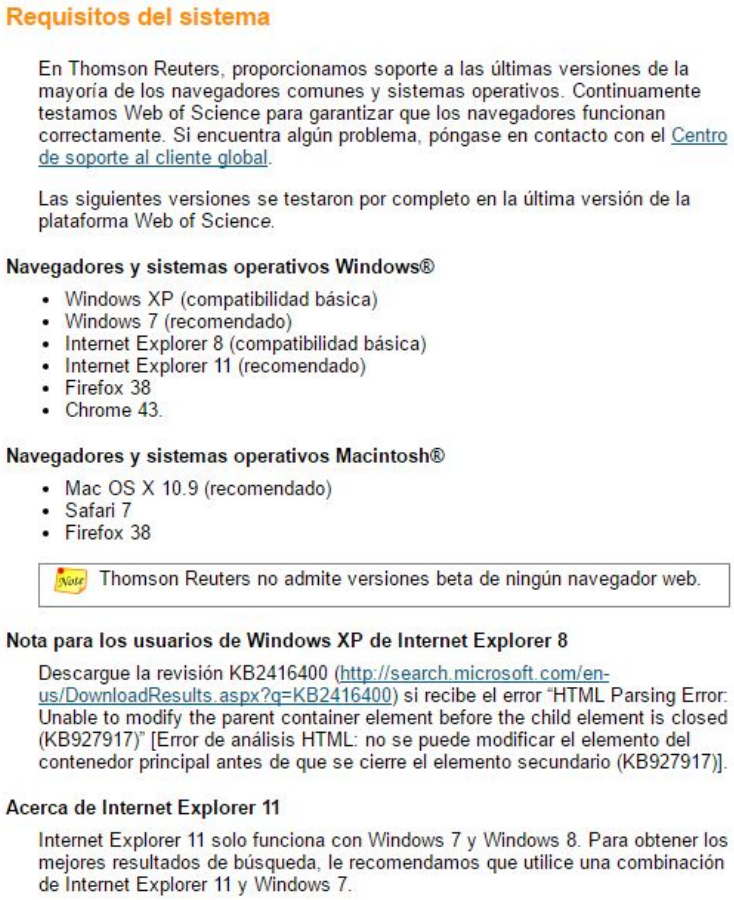


Figura 2.25: h-*graph* resultante de una búsqueda por autor en *Scopus*

Estas visualizaciones se encuentran disponible desde el sitio web de *Scholarometer* [59]. Son visualizaciones realizadas con el *plugin Adobe Flash Player*, habiendo aclarado esto, las conclusiones parecen ser obvias, no todos los navegadores ni todos los sistemas operativos son capaces de reproducir este tipo de representaciones gráficas. Es más, se comprobó en este trabajo que la visualización *Author Network* no funciona en navegadores actualizados (*Google Chrome* 54.0.2840.99 m y *FireFox* 47.0.2), el filtro por autor "no logra filtrar" los autores visualizados ni permite filtrar por autores distintos a los mostrados. No solo eso, como se mencionó en los apartados iniciales, *Scholarometer* no posee una interfaz web de consulta sino que lo hace a través de extensiones del navegador para *Chrome* y *Firefox*, una vez instalada la extensión, el navegador puede ejecutar consultas al motor de esta herramienta. Sin embargo, al igual que con la *WoS*, se sospecha que esta herramienta no está preparada para versiones nuevas de navegadores, ya que la interfaz de consulta no funciona, como se observa en las Figuras 2.29 y 2.30.

Por ello no se muestran gráficas de los resultados de esta herramienta por la imposibilidad de poder obtenerlos. De este modo se concluye que es una gran falencia de todos estos motores de búsqueda y herramientas académicas, el no contar con visualizaciones acordes y adaptables a los tiempos que corren, seguramente con el empleo de librerías compatibles

[59] http://scholarometer.indiana.edu/explore.html



Requisitos del sistema

En Thomson Reuters, proporcionamos soporte a las últimas versiones de la mayoría de los navegadores comunes y sistemas operativos. Continuamente testamos Web of Science para garantizar que los navegadores funcionan correctamente. Si encuentra algún problema, póngase en contacto con el Centro de soporte al cliente global.

Las siguientes versiones se testaron por completo en la última versión de la plataforma Web of Science.

Navegadores y sistemas operativos Windows®

- Windows XP (compatibilidad básica)
- Windows 7 (recomendado)
- Internet Explorer 8 (compatibilidad básica)
- Internet Explorer 11 (recomendado)
- Firefox 38
- Chrome 43.

Navegadores y sistemas operativos Macintosh®

- Mac OS X 10.9 (recomendado)
- Safari 7
- Firefox 38

Note Thomson Reuters no admite versiones beta de ningún navegador web.

Nota para los usuarios de Windows XP de Internet Explorer 8

Descargue la revisión KB2416400 (http://search.microsoft.com/en-us/DownloadResults.aspx?q=KB2416400) si recibe el error "HTML Parsing Error: Unable to modify the parent container element before the child element is closed (KB927917)" [Error de análisis HTML: no se puede modificar el elemento del contenedor principal antes de que se cierre el elemento secundario (KB927917)].

Acerca de Internet Explorer 11

Internet Explorer 11 solo funciona con Windows 7 y Windows 8. Para obtener los mejores resultados de búsqueda, le recomendamos que utilice una combinación de Internet Explorer 11 y Windows 7.

Figura 2.26: Requisitos del sistema que impone la *WoS* para visualizar el *Citation Map*

en múltiples navegadores o mediante la utilización de estándares como *JavaScript*, *CCS3* y/o *HTML5* se puedan lograr resultados interesantes y complementar los resultados tabulares que se entregan.

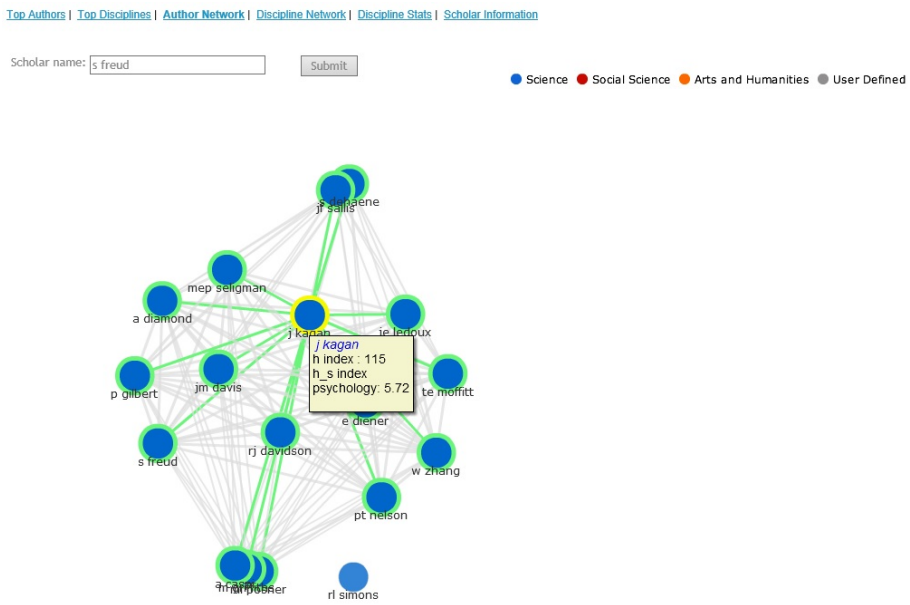


Figura 2.27: *Author Network* en *Scholarometer*

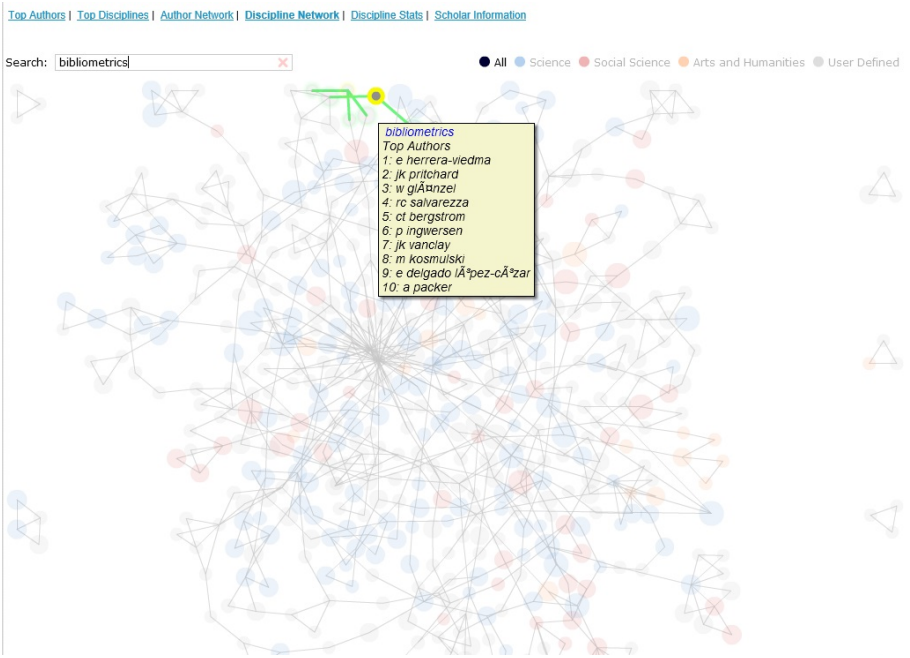


Figura 2.28: *Discipline Network* en *Scholarometer*

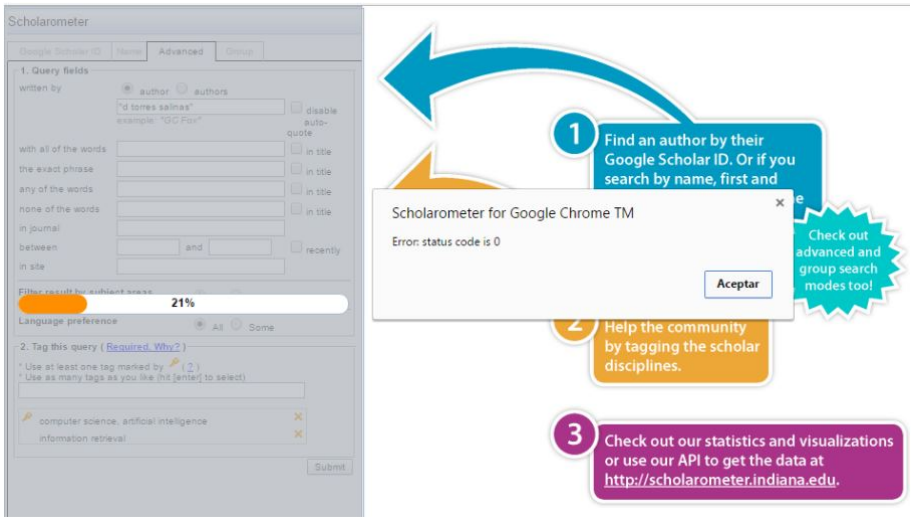


Figura 2.29: Búsqueda avanzada en *Scholarometer* utilizando *Google Chrome*

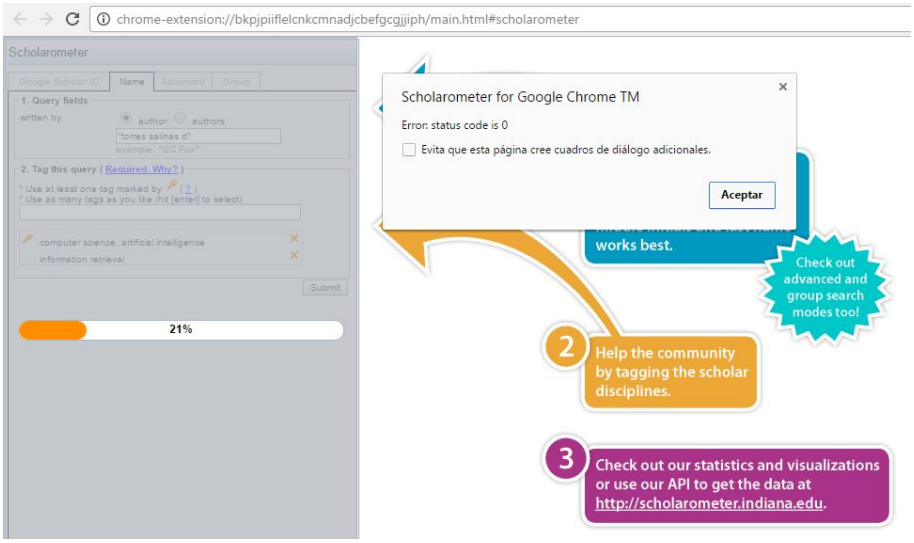


Figura 2.30: Búsqueda por nombre en *Scholarometer* utilizando *Google Chrome*

Capítulo 3

Metodología propuesta

3.1. Introducción

Para el desarrollo de este trabajo se emplearon una serie de herramientas, librerías y distintas tecnología, las cuales permitieron integrar en un todo, una solución aproximada pero válida a los problemas derivados de la falta de normalización en los registros de publicaciones indexados por motores de libre acceso. Si bien como se detallará más adelante, la herramienta funciona para cualquier fuente de datos sea esta de un motor académico de libre acceso o no, el único requerimiento es el formato de los datos a importar, con lo cual y sólo en teoría, la herramienta construida proveerá resultados significativos para datos de fuentes de motores de libre acceso pues en dichos motores es casi una regla general la falta de normalización, por el contrario, los datos que provienen de motores o bases de datos de pago, el contenido y la normalización de los registros de publicaciones están mucho más controlados, en este sentido la herramienta no entregaría resultados muy significativos, pero ello no indica que en este tipo de bases de datos no existan problemas de normalización como se mencionaron en apartados anteriores.

3.2. Diseño del *crawler*

Para dar cumplimiento a los objetivos planteados en esta tesis, se desarrolló una herramienta llamada "Academic Evaluator (AE)", la mis-

ma pretende e intenta en gran medida dar una solución aproximada a varios de los problemas derivados de la indexación de motores académicos, sobre todo motores de libre acceso como es el caso de *Microsoft Academic* y *Google Scholar*. De todas formas la herramienta funciona para cualquier origen de datos que respete el formato de entrada y que se detallará en apartados posteriores. Al ser una herramienta de evaluación de datos, estos datos deben ser recolectados de alguna forma, ya sea mediante un *crawler* o importando los datos recolectados desde cualquier origen. Un *crawler* es una herramienta básica para recopilar información estructural de un sitio web. También conocido como *spider* (araña), es un agente de software capaz de atravesar la web, comienza visitando una URL de una lista inicial, llamada semilla. A medida que el rastreador visita estas URLs, identifica el código HTML (los hipervínculos en la mayoría de los casos) que se necesita en la tarea de recuperación. La forma más común de identificar estas porciones de código dentro del documento recuperado, es mediante expresiones regulares. Para este trabajo se desarrollaron dos *crawlers* muy distintos pero con el mismo propósito, recuperar los datos de publicaciones científico-académicas que tanto *Google Scholar* como la API de *Microsoft Academic* entregan al realizar una consulta por autor.

3.2.1. Diseño del *crawler* de GS

En un principio la herramienta se pensó para que realizara la recuperación de información, tomando como fuente de datos GS, de forma automática, para ello se diseñó e implementó un *crawler* basado en *web scraping*. Se utilizó este tipo de mecanismo de recuperación debido a que a día de hoy GS no posee una API para poder recuperar datos de forma automática o al menos semiautomática.

El *crawler* construido posee un funcionamiento sencillo, realiza las consultas mediante peticiones HTTP *request* a la URL que entrega GS al momento de hacer las consultas, por ejemplo:

```
https://scholar.google.com/scholar?as_sdt=1,5&q=autor:daniel+
autor:torres+autor:salinas&hl=es&oe=ASCII&num=20&as_vis=1
```

Al ingresar la URL anterior en un navegador, se obtiene como resultado lo siguiente (ver Figura 3.1):

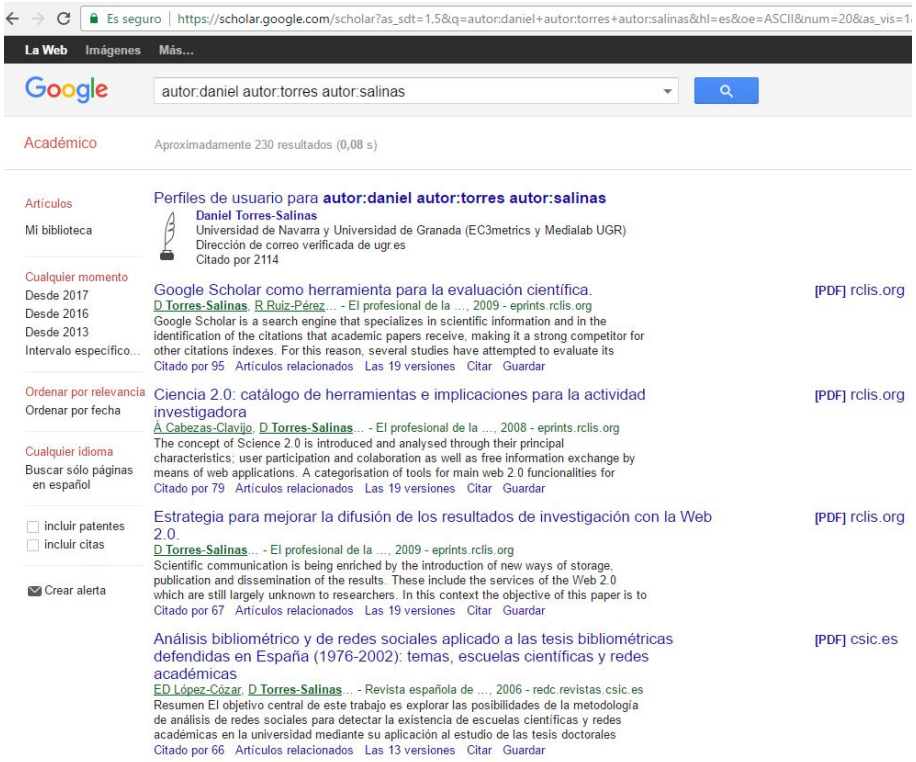


Figura 3.1: Página de resultados al realizar una búsqueda por autor en GS

Como se advierte en la figura anterior, no se tienen en cuenta registros que sean patentes o citas, por ello las casillas "incluir patentes" e "incluir citas" se encuentran desactivadas. Estas mismas opciones aparecen en la URL antes descrita, con lo cual se facilita la no inclusión de estos tipos de resultados. La razón de incluir o de no incluir estos registros es porque en este trabajo sólo se examinarán registros que respondan a publicaciones realizadas por el autor objeto de la búsqueda.

Una vez realizada la primera consulta, mediante el uso de expresiones regulares se van analizando los patrones que corresponden para la detección de un registro de *Google Scholar*. Del mismo modo que (Orduña-Malea et al., 2016a), aquí se describen las partes de un registro recolectado de *Google Scholar*(ver Figura 3.2):

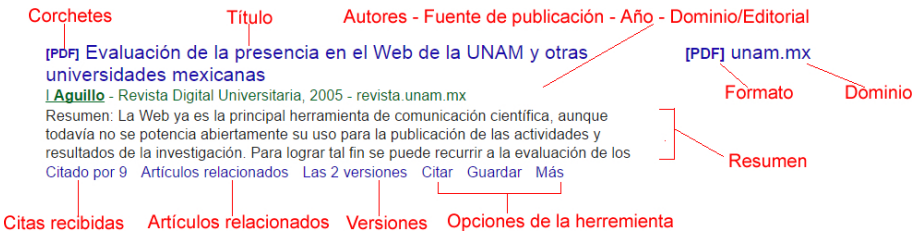


Figura 3.2: Elementos de un registro bibliográfico en *Google Scholar*

Corchetes : En este campo se indica la tipología documental del registro en cuestión, que aparece entre corchetes justo delante del título. Los elementos más comunes son los siguientes: PDF, HTML, BOOK o LIBRO y CITATIONS o CITAS.

Título : Se muestra el título de la publicación. GS ofrece en este elemento un link al sitio web donde el registro ha sido localizado. Este lugar no tiene por qué guardar concordancia con el campo "Dominio" del texto completo, pues la información bibliográfica del registro y el texto completo pueden estar en lugares diferentes.

Autores : Se muestra la lista de autores del documento separados por "; ".

Fuente de publicación : Aparece el nombre de la fuente donde el documento ha sido publicado y, en algunos casos, información descriptiva (volumen, número, paginación).

Año : Año de publicación, los registros que no poseen año de publicación se debe, en la mayor parte de los casos, a una indización incorrecta.

Dominio/Editorial : Aparece la URL donde el documento ha sido localizado. Éste no tiene por qué corresponder con el campo "Dominio" del texto completo. Adicionalmente, y sólo para las grandes editoriales, puede aparecer el nombre de la editorial (por ejemplo, Elsevier).

Formato : Este elemento junto con **Dominio** solamente aparecen en la parte derecha de la pantalla cuando *Google Scholar* dispone de una versión a texto completo de la versión del registro analizada, ambos campos se representan por un link al recurso (Martín-Martín et al.,

2016d). Se indica entre corchetes el formato del fichero en el que está disponible el texto completo.

Dominio : Dominio web en el que *Google Scholar* ha localizado la versión del documento a texto completo.

Resumen : Se muestran las primeras líneas del resumen del documento, este campo posee un límite, con lo cual no se muestra el resumen completo.

Citas recibidas : Se muestra el número de citas que el documento ha recibido según *Google Scholar*. Este campo posee un link al listado de las fuentes que citan la publicación.

Artículos relacionados : Enlace que dirige al usuario a un listado limitado de documentos que *Google Scholar* ha considerado similares o relacionados con el registro consultado y, por tanto, de posible interés para el usuario.

Versiones : Si un mismo documento es publicado en diferentes sitios web o repositorios, aquí se ofrece un link al conjunto de versiones de este documento.

Opciones de la herramienta : Adicionalmente, *Google Scholar* ofrece las opciones de *Citar* la publicación y descargarla en varios formatos, además de *Guardar* en la biblioteca del perfil del usuario.

Es preciso identificar la porción y el código HTML de cada registro para poder extraer los datos de forma correcta. Para este trabajo se descartan en una primera instancia los registros que no poseen año de publicación, esto es así pues el análisis que se realizará luego es un análisis de la productividad a través del tiempo, con lo cual es necesario e imprescindible conocer el año de publicación de un registro. Los registros que no poseen año de publicación corresponden principalmente a una indexación incorrecta o errónea, dado que resulta impensable que un "artículo real", escrito por "autores reales", publicado en algún "sitio real" no posea año de publicación. Si bien uno estaría tentado a pensar que, ya que se está utilizando *scraping* para recuperar los registros de GS y los registros "malformados" de GS poseen un enlace, en teoría, válido al recurso que está referenciando, se podría emplear *scraping* para averiguar el año de publicación del artículo desde dicho enlace. Bueno, esto

no es posible, al menos por el momento y al menos en lo que respecta al alcance de este trabajo ya que se deberían de construir *crawlers* para cada tipo posible de repositorio donde estos registros son almacenados, para poder detectar el código HTML que hace referencia al año de publicación, como esto es una tarea titánica por no decir imposible, ya que al no poseer un protocolo estándar, es imposible que un nuevo repositorio, revista o sitio web cumpla y respete un esquema aún no establecido de forma global para almacenar los metadatos de una publicación y poder establecer con exactitud un dato tan sensible como el año de publicación. Aquí cobran especial relevancia iniciativas como el protocolo OAI-PMH para la recolección de información y la importancia de la calidad de los metadatos (Medrano et al., 2012b), porque de nada serviría poseer un repositorio bien estructurado si los metadatos no respetan ciertas normas para facilitar la interoperabilidad entre estos, ejemplo de ello es la aplicación de las directrices DRIVER 2.0(Digital Repository Infrastructure Vision for European Research, 2008) u otras basadas en esta (OpenAIRE(OpenAIRE, 2013) y SNRD(Ministerio de Ciencia Tecnología e Innovación Productiva and Consejo Interinstitucional de Ciencia y Tecnología, 2013)) cuyo objetivo es la normalización de la representación de algunos metadatos y el cumplimiento de ciertos metadatos de forma obligatoria, recomendada u opcional, esta estandarización en la representación y codificación de metadatos, además de los protocolos de comunicación, permite la interoperabilidad entre los distintos sistemas de información. Por esta razón los registros recolectados que no poseen año de publicación son descartados en esta primera instancia.

Una vez detectado el año de publicación, se continúa analizando los demás datos de interés: Título de la publicación, Autores, Nombre de la revista o lugar de publicación, Año de publicación, Resumen, Tipo de archivo, Enlace al recurso, Enlace a las citas, Número de citas.

La estrategia de recolección continúa analizando las distintas páginas de resultados que entrega la consulta inicial (ver Figura 3.3):

Para poder recorrer estas páginas también se detecta mediante expresiones regulares el patrón del enlace, de este modo se puede saber cuándo se llegó a la última página de resultados.

Como se mencionó, la idea de la herramienta construida era realizar la recolección en línea, sin almacenar datos y llevar a cabo todo el proceso de análisis en tiempo real, con datos actualizados, sin embargo, en estos

Figura 3.3: Índice de páginas de resultados en *Google Scholar*

últimos meses *Google* ha puesto restricciones a los procesos automáticos de recolección de datos de cualquiera de sus motores, incluido el buscador *Google* de propósito general y lo más importante *Google Scholar*. Los robots de *Google* analizan ciertos patrones de comportamientos de los procesos de recolección detectando cuales parecen ser automáticos, en caso de detectarlos bloquean las consultas solicitando que el usuario resuelva un código CAPTCHA similar al mostrado en la Figura 3.4:

CAPTCHA (*Completely Automated Public Turing test to tell Computers and Humans Apart*) [1] son pruebas de *Turing* automatizadas utilizadas para determinar si el usuario final es humano y no un programa automatizado (robot). Se solicita a los usuarios que lean y respondan a un CAPTCHA visual, que a menudo se presenta como un mapa de bits de cadenas de caracteres, con el fin de obtener acceso a un recurso.

Una vez introducida la secuencia alfanumérica se verifica que sea correcta, y si esto sucede el sitio web entrega los resultados.

Justamente este bloqueo, la resolución del código CAPTCHA, fue lo que impidió que el *crawler* ya construido, funcionase de forma automática, limitando esto la recolección de los datos. Pues no se pudo resolver de forma automática dicho bloqueo, es más, a medida que uno realiza recuperaciones más exhaustivas o repetitivas los bloqueos son cada vez mayores, incluyendo en algunos casos la resolución del código CAPTCHA y la resolución de un conjunto de imágenes en las que se solicita se indiquen cuales poseen alguna característica (ver Figura 3.5).

Sin embargo no se debe perder de vista que la recolección automática de información de un autor es solo la parte inicial y una muy pequeña

[1] `http://www.captcha.net/`

Figura 3.4: Código CAPTCHA implementado por GS para recuperar resultados

del proyecto, pues el principal objetivo de este trabajo es la construcción de un esquema de desambiguación y visualización, la recolección de información, como se verá más adelante, es un proceso secundario y podrá realizarse desde cualquier herramienta o fuente de datos externo y luego ese conjunto de datos podrá ser importado a la aplicación para que sean procesados. De un modo u otro, la herramienta permitirá realizar la recolección automática de los datos utilizando como origen de los datos a *Microsoft Academic*, ya que esta herramienta si posee API para recuperar los resultados de las búsquedas.

Figura 3.5: CAPTCHA con imágenes implementado por GS para limitar la recuperación de resultados

Como el proyecto necesitaba los datos de GS primero para hacer pruebas y luego para comprobar el correcto funcionamiento del modelo presentado, ya que *Google Scholar* es por excelencia la base de datos menos normalizada, lo que se hizo fue utilizar el desarrollo anterior e implementarlo por separado en una aplicación de escritorio (*Windows Form* con *Visual Studio* 2015 bajo C#) utilizando para ello el objeto

WebBrowser (ver Algoritmo 3.1), el cual permite asignarle una URL al método *Navigate* para poder navegar el contenido de la misma.

```
WebBrowser.Navigate(URL);
```

Algoritmos 3.1 WebBrowser Object

Una vez hecho esto, el objeto *WebBrowser* visualiza el contenido de la URL solicitada mostrando el contenido de la misma, así mismo, este objeto posee una propiedad llamada *DocumentText* que permite recuperar todo el código HTML del objeto, es decir, el código HTML de la URL solicitada al mismo. Una vez obtenido este código se procede de la misma forma que en el caso de la recuperación automática. La recuperación de esta forma tampoco es del todo automática, pues es forzoso realizarla con ciertas precauciones como hacerla pausada, una consulta nueva cada 6″ o 10″ y esperar unos minutos entre autor y autor, pues para realizar las pruebas se recuperaron los registros de al menos 10 autores distintos. No obstante, tomando estas precauciones, los robots de *Google* siguen bloqueando este proceso, solicitando la resolución del código CAPTCHA y la secuencia de imágenes para descartar procesos automáticos.

La única diferencia en esta nueva implementación del *crawler*, es que al solicitar el código CAPTCHA se lo puede introducir y resolver de forma manual, continuando con la navegación. El proceso continúa navegando cada una de las páginas de resultado de la consulta inicial y procesando una a una dichas páginas para extraer las publicaciones del autor solicitado.

Un dato que se almacena, en caso de que exista, es la URL de las citas a esa publicación. Una vez que se terminan de procesar las páginas de resultados, se procesan las páginas de citas del mismo modo, esto se hace para poder desambiguar las citas a publicaciones o registros duplicados, este proceso tiene el nombre de fusionado de citas. Esto permitirá luego al desambiguador, tener una número de citas más cercano a la realidad, debido a que no sería correcto arbitrariamente decidir entre dos documentos que al parecer son duplicados, optar por el de mayor números de citas, puesto que se ha comprobado que en algunos casos existe solapamiento de citas y en otros, algunas citas de los documentos descartados no son tenidas en cuenta y no son atribuidas a una cierta publicación.

En esta segunda vuelta (haciendo referencia a la recolección de citas

de una publicación), si el servicio de *Google Scholar* solicita la resolución del CAPTCHA, después de que este se valide, se continúa con la recolección.

3.2.2. Diseño del *crawler* de MA

Afortunadamente, como se había adelantado, la nueva versión del motor académico de *Microsoft* denominado *Microsoft Academic* (MA) posee una API muy novedosa. La versión anterior de este motor (MAS) también poseía una API, no tan conocida, para la cual era necesario que el responsable del producto evaluase una nueva solicitud para obtener un ID (Identificador) para poder utilizarla, es decir, el acceso a esta API no era concedido a cualquier persona. Pero en esta versión mejorada del motor académico, el acceso se solicita mediante una cuenta *Microsoft* válida o cuenta de *GitHub* o *LinkedIn*. En el portal *Microsoft Cognitive Services*[2] se ofrece acceso a un gran número de APIs para diferentes proyectos (búsqueda de imágenes, procesamiento de lenguaje natural, procesamiento y analítica visual de imágenes, sistema de recomendación, traductor y muchas otras), en este caso, se solicitó una suscripción para la *Academic Knowledge API*[3], como se ve en la Figura 3.6, esta suscripción entrega al usuario que se suscribe 2 *Keys* (claves), indica la cantidad máxima de transacciones según el tipo de servicio que se utilice, para este caso se utilizará el método *Evaluate*[4], además permite regenerar esta clave como un mecanismo de seguridad y por último mostrar la cuota (opción *show cuota*) de utilización del servicio.

Estas limitaciones vienen dadas por ser un servicio gratuito, evitando así la sobrecarga innecesaria de los servidores de *Microsoft*, pero esto no limita la velocidad en la respuesta y la calidad de los datos recuperados.

La *Academic Knowledge API* (AK API) habilita a los usuarios a recuperar información de *Microsoft Academic Graph* (MAG). MAG es una base de datos que modela "las actividades de comunicación académica de la vida real como un grafo heterogéneo que consta de seis tipos de entidades" (Sinha et al., 2015). Estas entidades son: las publicaciones, el campo de estudio, el autor, la institución (afiliación del autor), el lugar (revista

[2] https://azure.microsoft.com/es-es/services/cognitive-services/

[3] https://tinyurl.com/zaranm9

[4] https://docs.microsoft.com/nl-nl/azure///cognitive-services/academic-knowledge/paperentityattributes

o serie de conferencias), y el evento (instancias de la conferencia). Los datos para MAG se recogen principalmente de los *feeds* de metadatos de editores y páginas web indexadas por *Bing*[5].

La AK API ofrece los métodos *Interpret*, Evaluate y *CalcHistogram* para recuperar datos de MAG. Los dos últimos son esenciales para los trabajos bibliométricos. El método *Evaluate* recupera un conjunto de atributos basados en una expresión de consulta. Las expresiones de consulta pueden construirse con atributos de entidad. Una solicitud *Evaluate* produce uno o varios resultados coincidentes, o ninguno, en caso de que no haya coincidencia. Cada resultado contiene un valor de probabilidad de registro natural para indicar la calidad de la coincidencia. Por lo tanto, el método *Evaluate* es un medio para recopilar metadatos sin procesar de MA. Por el contrario, el método *CalcHistogram* calcula un histograma de la distribución de valores de atributo para las entidades devueltas por una expresión de consulta, tal como la distribución de citas por año para un autor determinado.

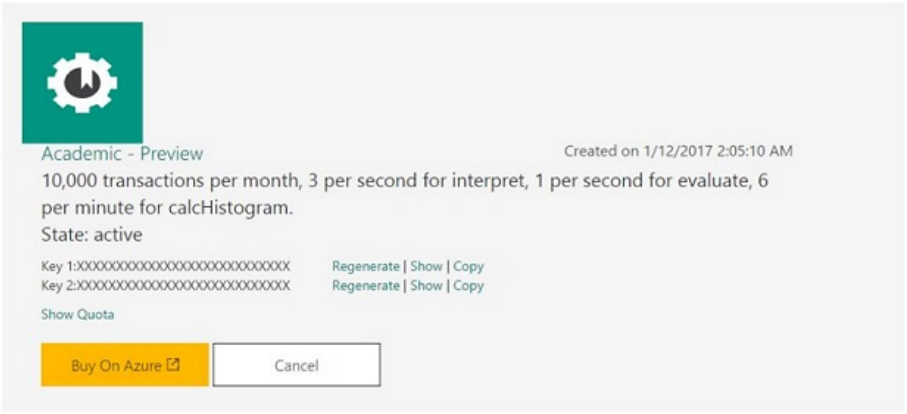


Figura 3.6: Restricciones para el uso de las claves de la *Academic Knowledge API*

En la AK API, existen 20 atributos de entidad que se pueden utilizar para generar expresiones de consulta, así como para especificar la respuesta a una consulta. Nueve atributos están vinculados al documento de la entidad (*paper*, artículo), cinco al autor de la entidad y dos a cada

[5] https://www.bing.com/

una de las entidades campo de estudio, revista (*journal*) y lugar (*venue*):

Artículo : ID, título, lenguaje de la publicación, año de publicación, fecha de publicación, el número de citas, el número de citas estimado, el ID de referencia y las palabras incluidas en el título y el resumen.

Autor : nombre del autor, ID de autor, afiliación, ID de afiliación, orden del autor en la publicación.

Campo de estudio/revista/lugar : nombre, ID.

Además, hay 13 atributos de metadatos ampliados que, a diferencia de los 20 atributos de entidades, sólo pueden utilizarse para especificar la respuesta de la consulta. Los 13 atributos de metadatos extendidos están disponibles para las entidades *artículo* (once atributos) y el *lugar* (dos atributos):

Artículo : nombre de presentación del documento, URL de origen, formato (por ejemplo, HTML, PDF, PPT), volumen, edición, primera página, última página, DOI, Contextos de citas, Número de elementos del índice (recuento de palabras del resumen) y Lista de palabras abstractas y su posición correspondiente en el resumen original.

Lugar : nombre para mostrar, nombre abreviado

El detalle de estos se encuentra indicado en (Services, 2017), allí se puede examinar el conjunto completo de atributos y metadatos asociado a una publicación o registro y cuales son los tipos de datos de cada uno para poder recuperarlos de forma correcta. Resulta útil conocer el tipo de datos de los atributos de las entidades, pues en algunos casos (los metadatos sobre todo), vienen representados como un *array* de *array*.

El procedimiento de recolección es sencillo, para solicitar los registros de un autor determinado se realiza una consulta HTTP como la siguiente (ver Algoritmo 3.2):

```
var client = new HttpClient();
var queryString =
    HttpUtility.ParseQueryString(string.Empty);
```

```
// Request headers
client.DefaultRequestHeaders.Add("Ocp-Apim-Subscription
-Key", "XXXXXXXXXXXXXXXXXXXX31a5e25cc838");

// Request parameters
queryString["expr"] = "Composite(AA.AuN=='"+
    szAuthorName + "')";
queryString["model"] = "latest";
queryString["count"] = "1000";
queryString["offset"] = "0";
queryString["orderby"] = "Y:desc";
queryString["attributes"] =
    "Id,Ti,Y,CC,AA.AuN,AA.AuId,AA.AfN,


var uri =
    "https://westus.api.cognitive.microsoft.com/academic/
v1.0/evaluate?" + queryString;
```

Algoritmos 3.2 Parámetros necesarios de Academic Knowledge API

En la sección *Request headers* se indica la *Key* obtenida al realizar la suscripción al servicio. En la sección *Request parameters* se fijan todos los parámetros de la consulta, el parámetro más importante es *expr* que servirá para indicar el nombre del autor que se está buscando (para el caso de la recuperación de información realizada en este proyecto), este parámetro puede incluir otras entidades además del nombre de autor como ser: Revista, Serie de Conferencia, Afiliación, Campo de Estudio. Otro atributo parametrizable es *attributes*, en este atributo se pueden indicar todos los atributos que sean necesarios recuperar, el resto de parámetros (inclusive *attributes*) son opcionales; en *model* se indica el nombre del modelo que desea consultar, actualmente, el valor predeterminado es *latest*; *count* se utiliza para señalar la cantidad de resultados a recuperar, para este trabajo se estableció en 1000 (solo porque GS recupera los primeros 1000 registros); *offset* es el índice del primer resultado a devolver; y *orderby* es el nombre del atributo que será utilizado para ordenar de forma ascendente o descendente las entidades.

La respuesta del motor siempre viene en formato JSON (JavaScript Object Notation)[6] con tres atributos: *expr* (la cadena indicada como *expr* en el *request*), *aborted* (con valor true en caso de que el *request*

[6] http://www.json.org/

haya superado el tiempo máximo de respuesta) y por último *entities* (es un *array* de 0 o más entidades que coinciden con los parámetros de la búsqueda).

Una vez realizada la consulta (ver Algoritmo 3.3):

```
var response = await client.GetAsync(uri);
var jsonResponse = await
   response.Content.ReadAsStringAsync();
dynamic data = JObject.Parse(jsonResponse);
```

Algoritmos 3.3 Consulta y respuesta a Academic Knowledge API

Se recorre el resultado para procesar las entidades (ver Algoritmo 3.4):

```
foreach (dynamic rec in data.entities)
{
  Structures.STRecord oRecord = new
     Structures.STRecord();
  oRecord.CiteNumber = rec.CC;
  oRecord.Patron = szAuthorName;
  oRecord.Source = "MA";
  oRecord.Year = rec.Y;
  oRecord.URLCites =
     "https://academic.microsoft.com/#/detail/"
                    + rec.Id;
  oRecord.JournalURL = rec.JN;
  foreach (dynamic aut in rec.AA)
  {
      oRecord.Authors += aut.AuN + ","; //aut.AuId
  }
  oRecord.Authors = oRecord.Authors.Substring(0,
                    oRecord.Authors.Length - 1);
  string szE = rec.E;
  dynamic json2 = Json.Decode(@szE);

  if (json2.DN != null)
     oRecord.Title = json2.DN;
  else
     oRecord.Title = rec.Ti;

  if(json2.D!=null)
     oRecord.Abstract = json2.D;
```

```
  if (json2.S != null)
  {
      foreach (dynamic dato in json2.S)
      {
          switch ((int)dato.Ty)
          {
              case 1:
                  oRecord.Type = "[HTML]";
                  break;
              case 2:
                  oRecord.Type = "[TXT]";
                  break;
              case 3:
                  oRecord.Type = "[PDF]";
                  break;
              case 4:
                  oRecord.Type = "[DOC]";
                  break;
              case 5:
                  oRecord.Type = "[PPT]";
                  break;
              case 6:
                  oRecord.Type = "[XLS]";
                  break;
              case 7:
                  oRecord.Type = "[PS]";
                  break;
              default:
                  oRecord.Type = "";
                  break;
          }
          oRecord.URL = dato.U;
          break;
      }
  }
  lstRecords.Add(oRecord)
}
```

Algoritmos 3.4 Recuperar registros de la consulta a AK API

La porción de código anterior recorre el *array* de entidades entregadas como resultado y las almacena en una lista para luego guardar los resultados en una tabla que tendrá todos los registros en crudo antes del

procesamiento.

Algo que se echa en falta y resulta un tanto limitante, es que la API no posee un listado de las publicaciones que citan la entidad que se está procesando, la respuesta retorna la cantidad de citas recibidas pero solo es un número. Si bien se puede armar la URL para acceder a la página donde aparecen listadas las publicaciones que citan la publicación en cuestión, no es posible procesarla mediante *scraping*, al parecer *Microsoft* ha implementado algún mecanismo de control/seguridad para impedir ver el código fuente de una página de resultados. Por esta razón, los registros recuperados de MA solo poseen el número que indica la cantidad de citas y a diferencia de los registros recuperados con GS (u otra fuente que si lo permita), no poseen el conjunto de publicaciones que citan el mencionado trabajo.

Sin embargo, la facilidad en la recuperación de información así como la riqueza de los metadatos entregados en contraposición con lo laborioso que resulta recuperar registros y metadatos en GS (los únicos metadatos provistos por GS son: identificador, autores, título, origen, año, volumen, edición, páginas, editores y número de citas), se concluye al igual que (Hug et al., 2016), que MA supera a GS en términos de funcionalidad, estructura y riqueza de los datos, así como en lo que respecta a la recuperación de datos y manejo.

3.3. Diseño del esquema de desambiguación

En los apartados anteriores dedicados a la revisión bibliográfica, se pudo ver y en cierta forma tomar noción de la complejidad de los problemas derivados de la ambigüedad a la hora de cuantificar la productividad de un investigador, tomando como fuente el conjunto de publicaciones científico-académicas. Si el origen de datos es una base de datos poco controlada o con mecanismos de indización poco eficientes, como es el caso de los motores de libre acceso (que es el objeto de comparación en este trabajo) los problemas aumentan de manera muy importante.

Algunos de los problemas derivados de estos mecanismos de indización son:

- publicaciones duplicadas incrementando así los indicadores basados en la cantidad de publicaciones;

- recuento de citas incorrecto de un algún trabajo a un autor por firmar de forma distinta entre publicaciones, por errores ortotipográficos en el nombre, por la existencia de homónimos;
- por la existencia de publicaciones duplicadas y por la posibilidad de referenciar a cualesquiera de estas, se produce una partición del conjunto de citas que recibe una publicación entre estas dos publicaciones que para el motor parecen ser distintas (ver Figura 3.7).

(a) Publicación con 11 citas

(b) Publicación con 12 citas

Figura 3.7: Publicación duplicada en *Google Scholar* con conjunto de citas distintas para cada una

En la extensa bibliografía existen varias y variadas aproximaciones a cada uno de los problemas detectados, y si uno advierte no es el mismo problema, sino que son variaciones dependiendo del origen de este. Es más, las soluciones que existen o que algunos autores han intentado dar, se reducen a atacar un solo problema a la vez, en un conjunto bien definido, controlado y acotado de datos (los investigadores en Ciencias Naturales del Reino Unido por dar un ejemplo) o simplemente detectan la existencia del problema y se limitan a informarlo. Claro está, que con

la utilización obligada de estándares como DOI y ORCID, muchos de estos problemas no existirían, pero también es claro que el uso de estas herramientas no es masivo y no son del todo gratuitas (caso de DOI).

Este trabajo puede parecer un tanto ambicioso, pero el mismo intenta dar un paso más allá de las soluciones existentes, ya que pretende brindar una solución aproximada a la gran parte de estos problemas analizando la información de publicaciones recuperadas de alguna base de datos académica (para este trabajo, los registros de *Google Scholar* por un lado y *Microsoft Academic* por el otro). Para lograr esta aproximación se desarrolló un conjunto de algoritmos para atacar por partes ("divide y vencerás") las variantes de los problemas mencionados, los cuales se detallarán a continuación y que se pueden resumir en la Figura 3.8.

Figura 3.8: Esquema general del procedimiento a emplear

3.3.1. Algoritmo ágil, desambiguación de autores

El funcionamiento del algoritmo principal se basa en un conjunto de reglas lógicas, heurísticas y algunos supuestos. En una primera instancia es necesario recolectar la información a procesar, ya sea de forma automática desde la herramienta construida utilizando la AK API de MA o importando los datos de una fuente externa. Se podría hacer la recuperación automática desde *Google Scholar* pero no se asegura el correcto funcionamiento debido a los bloqueos mencionados en apartados anteriores. Para el caso de la recolección desde *Microsoft Academic* ya se indicó cómo se hace y qué parámetros incluir. Para el caso de la importación de datos de una fuente externa, el archivo a importar debe ser un archivo en formato JSON con la siguiente estructura:

```
{
```

``Title'': ``Scientometrics 2.0: New metrics of scholarly impact on
       the social Web'',
``Authors'': ``J Priem, BH Hemminger'',
%``Patron'': ``jason priem'',
``Year'': 2010,
``URL'': ``http://pear.accc.uic.edu/ojs/index.php/fm/article/
       viewArticle/2874'',
``URLCites'': ``http://scholar.google.com/scholar?cites=
           11984823944214187051&as_sdt=2005&
           sciodt=1,5&hl=es'',
``Type'': ``HTML'',
``CiteNumber'': 259,
``JournalURL'': ``pear.accc.uic.edu'',
``Abstract'': ``Abstract The growing flood of scholarly literature
           is exposing the weaknesses of current, citation-
          based methods of evaluating and filtering articles.
          A novel and promising approach is to examine the use
           and citation of articles in a new forum: Web 2.0
           services like social'',
``Source'': ``GS''
``Cites'':
 [
  { ``Title'': ``Can tweets predict citations? Metrics of social
       impact based on Twitter and correlation with traditional
         metrics of scientific impact'',
    ``Authors'': ``G Eysenbach'',
    ``Year'': 2011
    ``URL'': ``http://www.jmir.org/2011/4/e123/?utm_source=
         feedburner&utm_medium=feed&utm_campaign=
         Feed:+JMedInternetRes+''},
  { ``Title'': ``The digital scholar: How technology is transforming
         scholarly practice'',
    ``Authors'': ``M Weller'',
    ``Year'': 2011
    ``URL'': ``http://books.google.com/books?hl=es&lr=
         &id=Lj8lc8hWVvEC&oi=fnd&pg=PP1&ots=
         xL9tMlOSO4&sig=8_uE0JBRhsUO4iMWGNN6lGpcQAE''},
  { ``Title'': ``Do altmetrics work? Twitter and ten other social
         web services'',

```
  ``Authors'': ``M Thelwall, S Haustein, V Larivière, CR Sugimoto'',
   ``Year'': 2013
  ``URL'': ``http://journals.plos.org/plosone/article?id=10.1371/
        journal.pone.0064841''}
 ]
}
```

El detalle es el siguiente:

Title : (string[500]) Título de la publicación.

Authors : (string[500]) Lista de autores separados por "," (coma).

Year : (int) Año de publicación.

URL : (string[400]) URL a la publicación ya sea al documento o al repositorio donde está almacenado.

URLCites : (string[400]) URL al conjunto de citas de dicha publicación.

Type : (string[50]) Formato de la publicación, los formatos pueden ser: HTML, TXT, PDF, DOC, BOOK, PPT, XLS, PS u OTHERS.

CiteNumber : (int) Número de citas.

JournalURL : (string[250]) Para el caso de GS (de los datos que se recolectaron con el *crawler* implementado), dominio donde el documento ha sido localizado o en algunos casos el nombre de la Editorial, no se guarda el nombre de la revista pues gran parte de los registros no posee y en algunos casos aparece cortado. Para el caso de MA si se almacena el nombre de la revista, ya que la AK API si lo provee, aunque no todos los registros lo cumplimentan. La idea de este atributo es indicar un parámetro para agrupar un conjunto de registros, por consiguiente es indistinto si se utiliza el dominio, editorial o nombre de la revista, luego en la visualización se verá la utilidad del mismo.

Abstract : (string[3000]) Resumen (abstract) de la publicación.

Source : (string[5]) Abreviación del origen de los datos, por ejemplo: GS para *Google Scholar* o MA para *Microsot Academic*.

Cites : Es un *array* de objetos utilizado para almacenar las citas recibidas por la publicación en cuestión, si no es posible recuperar las citas (porque el motor académico no lo permite) o la publicación no posee citas, este atributo es vacío. Cada elemento posee los siguientes atributos: *Title, Authors, Year* y *URL*, los cuales poseen la misma especificación y significado que lo indicado para la publicación.

Una vez recolectados o importados los datos, estos se agruparán en primera instancia por el parámetro/campo *patrón* = *nombre* + *apellido*, a saber, mediante este atributo se realizará la búsqueda en los distintos registros almacenados que respondan a dicho *patrón*, si existiese más de una fuente de datos almacenada (parámetro/campo *source*), la aplicación permitirá elegir sobre cual fuente se realizará la comparación.

El algoritmo recorre uno a uno este conjunto de registros seleccionados (publicaciones) buscando que el *patrón* (que será el nombre del autor objeto de análisis) coincida o sea parecido a alguno de los *Autores* de la publicación examinada. Para llevar a cabo esta tarea se realizan las comprobaciones basándose en el análisis de *CoAutoría*, debido a que este esquema es el que mejor se adapta a la naturaleza del problema, ya que son datos que provienen de motores de libre acceso donde la normalización de datos no es el requisito principal, dónde pueden existir múltiples variantes para un mismo nombre y donde a priori no se conoce nada del autor o la información es escasa (por ejemplo no todas las fuentes retornan los datos de afiliación o campos de estudio), esto es, a diferencia de algoritmos basados en supervisión o aprendizaje automático, donde en la mayoría de los casos una red neuronal es entrenada con condiciones y casos de usos finitos y específicos, para este desarrollo se parte de la base del desconocimiento de cómo pueden variar no solo el nombre de un autor, sino también las relaciones con los coautores y con los autores homónimos. Entonces, siguiendo con la descripción del algoritmo, para cada registro se comprueba lo siguiente (ver Figura 3.9):

- Paso 1: Si el autor escribe solo (caso de único autor), los registros que coincidan exactamente con el nombre de autor buscado (**AB**) o con las variantes del nombre, se indicará que pertenece a dicho autor. Con lo cual, se establece que el registro actual pertenece al conjunto de publicaciones de este autor.

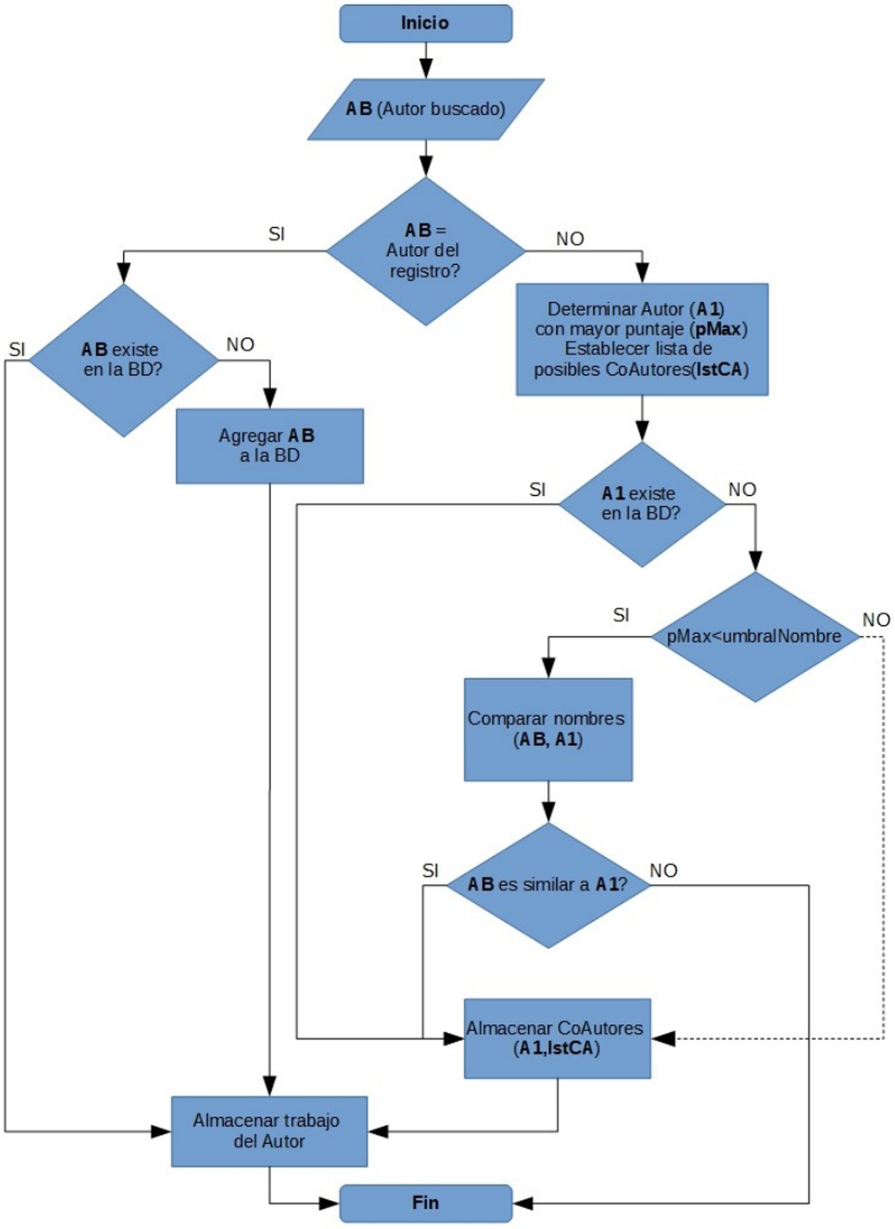


Figura 3.9: Diagrama de flujo del proceso de desambiguación

Cada Autor tendrá una tabla de publicaciones asociada a ese Au-

tor, la tabla se llama *Papers*, esta servirá para poder contabilizar las publicaciones de cada autor detectado. La búsqueda se realiza sobre la tabla *Authors* para los nombres de autor y en la tabla *VariantsAuthors* para las variantes del nombre de autor. En la tabla *VariantsAuthors* se almacenan las variantes encontradas según la indexación de la fuente de datos. Por ejemplo: al buscar los registros escritos por "Emilio Delgado Lopez Cozar", utilizando *Google Scholar*, existen al menos 16 formas distintas del nombre (como se puede apreciar en la Tabla 3.1). Claro está que antes de hacer cualquier comprobación y búsqueda, los nombres de autor deben ser normalizados internamente, por eso las tildes o acentos son eliminados, los caracteres ",", ";", "(", ")", "-", "_", "[", "]" y "." son reemplazados por el carácter de espacio en blanco(" "). Como se verá más adelante, en la tabla *Authors* pueden existir registros idénticos, autores con el mismo nombre, esto se debe a dos cuestiones que se tratarán en lo sucesivo: la existencia de homónimos y la diversidad de grupos de personas con las que un autor escribe.

Tabla: 3.1: Variantes del nombre Emilio Delgado López-Cózar encontradas en GS

Autor	Variantes del nombre en GS
Emilio Delgado Lopez Cozar	Delgado López-Cózar E
	Delgado-López-Cózar
	E Delgado López Cózar
	E Delgado Lopez-Cozar
	E Delgado López-Cozar
	E Delgado López-Cózar
	E Delgado-Lopez-Cozar
	E Delgado-López-Cózar
	E López-Cózar
	ED Lopez-Cozar
	ED Lopez-Cózar
	ED López-Cózar
	EL Cózar
	LCE Delgado
	DEDL Cózar
	EDL Cózar

- Paso 2: Para el restos de los casos, lo primero es identificar si en la lista de autores del registro a procesar, se encuentra incluido el autor objeto de estudio. Para ello se procesan uno a uno los nombres de autor (los distintos nombres vienen separados por coma ","), utilizando la función de similitud *MongeElkan*(Monge and Elkan, 1996; Cohen et al., 2013b), un esquema híbrido que se comporta muy bien para comparaciones basadas en pocos términos con mínimos errores. Esta función compara dos cadenas de caracteres y entrega un puntaje entre [0, 1]. El nombre de autor, del listado de autores de la publicación, que posea el puntaje máximo (**pMax**) de comparación se tomará como el "nombre tentativo" (**A1**, hasta aquí es sólo una suposición, pues si ningún nombre coincide con el autor buscado, cualquiera con un puntaje mayor a 0 valida la condición, por ello se aclara que se toma como tentativo sujeto a las verificaciones que se indican a continuación), el resto de nombres (si los hubiere) se tomarán como posibles CoAutores (**lstCA**).

  - Paso 2.1: Antes de continuar, con el posible autor identificado se realiza la misma comprobación que en el Paso 1, o sea, se verifica si el autor o alguna variante del nombre ya se encuentra almacenada en la Base de Datos. Si ya se encuentra se procede a analizar los CoAutores.
  - Paso 2.2: Si el nombre no se encuentra y el puntaje obtenido por la comparación (pMax) supera el umbral *umbralNombre*, se asume que el nombre es "bastante similar", con lo cual se procederá a analizar los CoAutores.
  - Paso 2.3: Si el nombre no se encuentra y el puntaje obtenido por la comparación no supera el umbral *umbralNombre*, se procede a realizar un conjunto de comprobaciones (*CompararNombres(AB, A1)*) con los términos del nombre que se indican en el Paso 3.
  - Paso 2.4: Si la comparación de nombres indica que ambos nombres son muy similares se procederá a analizar los CoAutores, caso contrario el registro se marca como "no comprobable". Estos registros marcados como "no comprobables" serán analizados en una segunda ronda, la explicación de ello se dará más adelante.

- Paso 3 (Comparación de Nombres): El proceso de comparación de nombres recibe como parámetros el conjunto de términos del nom-

bre del autor buscado (**AB** $\Rightarrow$ **S**) y el conjunto de términos del autor con el máximo puntaje (**A1** $\Rightarrow$ **T**). Se hace hincapié en que son conjuntos (*sets*) pues se aplicarán operaciones de conjuntos entre los términos. En una primera instancia se eliminan los términos en común de ambos conjuntos y se obtienen los elementos distintos (**DifST** que posee los elementos de **S** excepto los de **T** y **DifTS** que posee los elementos de **T** excepto los de **T**). Luego de esto se toma cada nuevo conjunto y se construye para cada uno dos listas (en total serían cuatro listas) de *strings* insertando en una los términos que se consideran iniciales (**lIniS**, por ejemplo: EDL que corresponde a Emilio Delgado Lopez) y en la otra los términos en sí mismos (*lTermS*). Luego de esto:

- Paso 3.1: Si solo existen términos en ambas listas, se concatenan los términos de cada una y se comparan mediante la función *MongeElkan*, si el resultado es mayor a *umbralNombre2* se concluye que existe una alta coincidencia, caso contrario se pasa al Paso 3.2.
- Paso 3.2: Ya sea que existan iniciales en algunas de las listas o que la comparación anterior no haya sido efectiva, esto quiere decir que existen iniciales o términos sin coincidir en su totalidad, por ello lo que se hará es intentar emparejar las iniciales que existen en **S** con los términos de **T**, dicho de otro modo, se comprobarán una a una las iniciales para ver si alguno de los términos restantes de **T** inician con dicha inicial, si es así, este término encontrado se elimina. Al finalizar esta comprobación quedarán los términos de **T** (**TermT**) y las iniciales de **S** (**IniS**) aún sin emparejar. De la misma forma se procede con las iniciales de **T** y los términos de **S**, obteniendo como resultado los términos de **S** (**TermS**) y las iniciales de **T** (**IniT**) aun sin emparejar, se pasa al Paso 3.3.
- Paso 3.3:
  - Situación 1: Si aún quedan términos en ambas listas, estos se concatenan por separado y se comparan mediante la función *MongeElkan*, si el resultado no es superior al valor *umbralNombre2* se concluye que ambos nombres no son similares, caso contrario, si no quedan iniciales en **S** (**lIniS** es vacío) se concluye que existe coincidencia entre ambos nombres ya que **S** es igual o menos específica

que **T**, con lo que si aún quedan iniciales en **T** no importa. Para el resto de los casos se concluye que no existen coincidencias.

- Situación 2: Si no quedan términos en ningunas de las listas, y no quedan iniciales en **S** (**lIniS** es vacío) se concluye que existe coincidencia entre ambos nombres ya que **S** es igual o menos específica que **T**, con lo que si aún quedan iniciales en **T** no importa. Para el resto de los casos se concluye que no existen coincidencias.
- Situación 3: Si no quedan términos ni iniciales en **S** se concluye que existe coincidencia entre ambos nombres. Para el resto de los casos se concluye que no existen coincidencias.

El diagrama de flujo de este proceso se encuentra descripto en la Figura 3.10.

- Paso 4 (Análisis de oautores): Según la situación que se determine, como se mencionó previamente, se deben analizar los coautores, este proceso se encargará de agregar a la base de datos el Autor y los CoAutores de la publicación en base a un conjunto de reglas. Como el esquema elegido para la desambiguación es el análisis de coautoría, resulta fundamental elegir los candidatos que formarán parte de las listas de CoAutores de cada Autor. Se pueden presentar las siguientes situaciones:

  - Situación 1: Si al autor escribe solo ya se contempló el procedimiento en el Paso 1.
  - Situación 2: En caso de que existan coautores, se consulta cada nombre de coautor en la base de datos por si existe el nombre (tabla *CoAuthors*) o alguna de sus múltiples variantes (tabla *VariantsCoAuthors*). Esta comparación es la clave del examen, ya que los nombres de autores de las publicaciones no siempre son indexados del mismo modo, por ello pueden existir variantes y combinaciones del nombre, desde el uso de iniciales hasta la omisión o agregado de nombres.

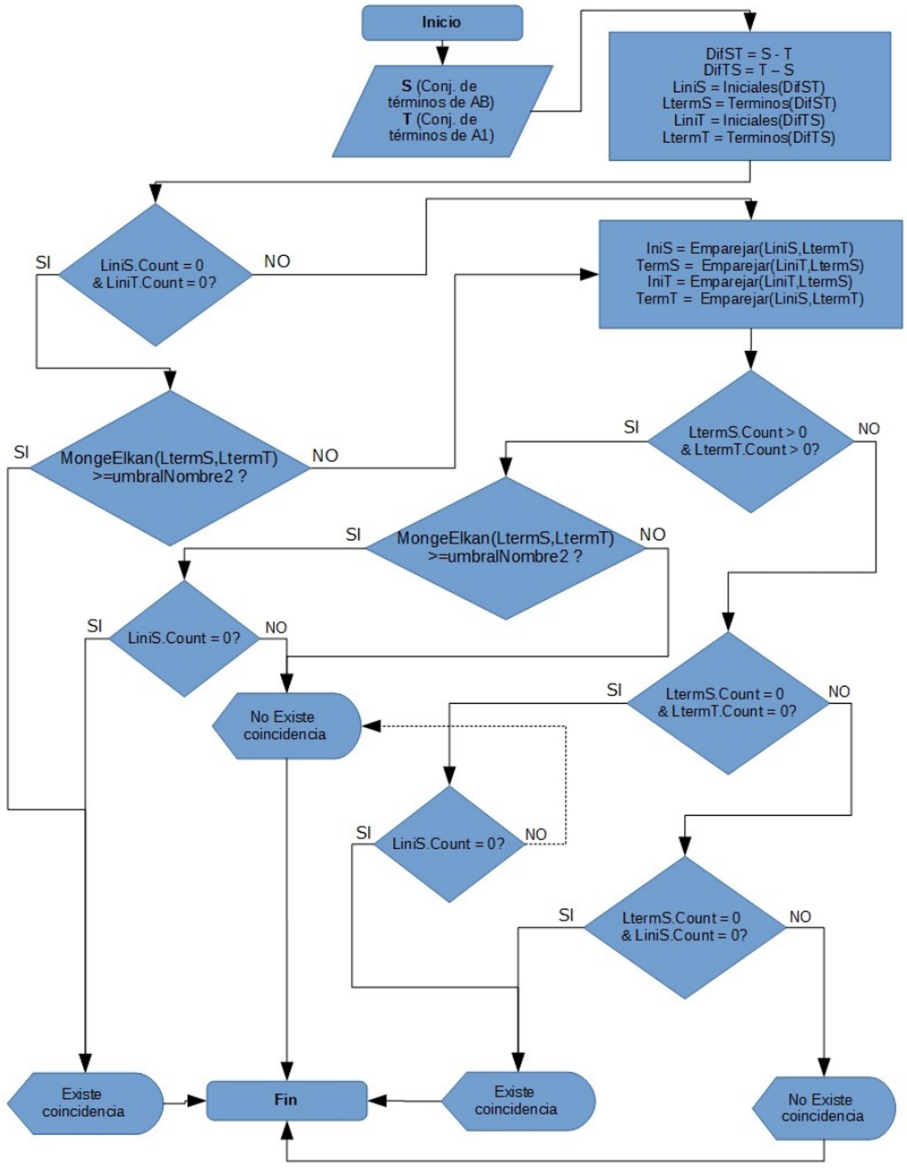


Figura 3.10: Diagra de Flujo del proceso de comparación de nombres

Aquí es necesario aclarar que las tablas que almacenan las variantes de los nombres de Autor y CoAutor no siguen la misma lógica, pues en la tabla *VariantsAuthors* se almacenan

sólo las variantes del nombre encontradas en las publicaciones que se determine que pertenezcan al autor en base a los CoAutores, por esta razón, para un mismo nombre de Autor pueden existir más de un registro en *Authors* pues la lista de CoAutores es distinta, son los CoAutores los que determinarán estas distintas agrupaciones. Por otro lado, en la tabla *VariantsCoAuthors* se almacenarán todas las posibles combinaciones y variantes realizadas por la aplicación, por ejemplo, para el autor *Emilio Delgado Lopez Cozar* las variantes serían las listadas en la Tabla 3.2

Tabla: 3.2: Combinaciones y variantes entregadas por la función *CombinacionesDeNombres(string t)*

Índice	Variantes del nombre
1	EDL Cozar
2	Lopez Cozar
3	Delgado Cozar
4	Delgado Lopez
5	Delgado Lopez Cozar
6	Emilio Cozar
7	Emilio Lopez
8	Emilio Lopez Cozar
9	Emilio Delgado
10	Emilio Delgado Cozar
11	Emilio Delgado Lopez
12	Emilio Delgado Lopez Cozar

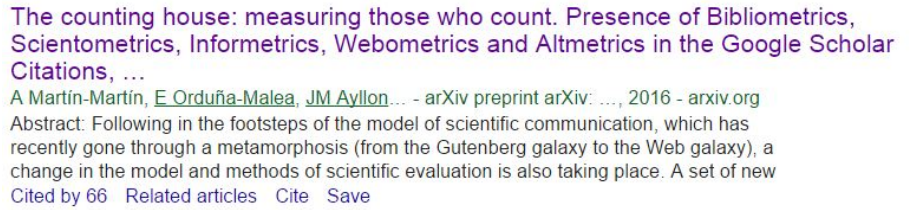
Al comparar los nombres de autores se tiene que considerar la especificidad del nombre de autor, esto es, cuanto más específico sea el nombre del autor, mejores resultados se obtendrán, primero en el proceso de recolección y luego en el proceso de análisis, en este sentido el nombre, de autor a recuperar y luego a analizar, no debe poseer iniciales (salvo que el autor sea conocido por las iniciales del nombre, ejemplo: "Isidro F. Aguillo") y debe buscarse por el nombre completo, luego los procedimientos de la aplicación desarrollada se encargarán de emparejar los nombres del autor con los términos almacenados de la publicación, pues en la mayoría de los casos, los nombres de autores de las publicaciones

son indexados con iniciales.

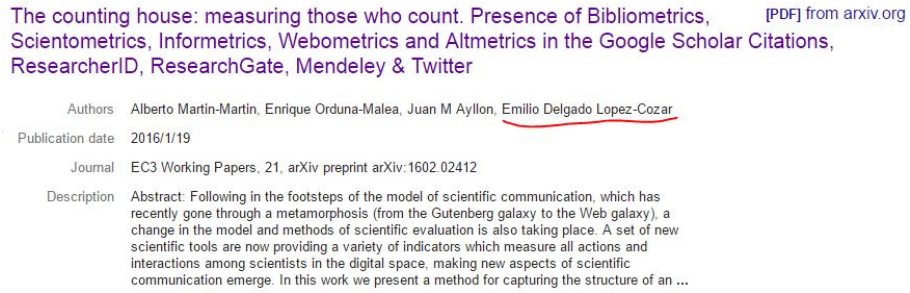
Para identificar las iniciales se toma como regla la sucesión de hasta tres letras seguidas en mayúscula, más de tres letras seguidas en mayúscula se toma como un término, es decir, un nombre de persona indexado en mayúscula. De igual modo, para extraer las iniciales de un nombre, se toma la primera letra de cada nombre siempre y cuando el nombre no sea un término detectado como una unión de iniciales en mayúscula, por ejemplo, para extraer las iniciales del nombre "ED Lopez Cozar" resulta en **EDLC**, pero para el nombre "JUAN LOPEZ", el resultado es **JL**, porque si bien todo está en mayúsculas, los términos poseen más de tres letra seguidas y son considerados nombres.

Una vez finalizada esta etapa inicial de procesamiento, algunos registros de publicaciones pueden no haber sido comprobados de forma satisfactoria, esto se debe a dos cuestiones, la primera de ellas indica que el autor buscado no se encuentra en el listado de autores de dicha publicación, la segunda indica que el autor buscado es más específico que algunos de los nombres encontrados en el listado de autores. Por ejemplo en la Figura 3.11a se presenta uno de los resultados de la búsqueda por autor en *Google Scholar* para "Emilio Delgado Lopez Cozar", como se advierte, aparecen tres autores: A Martín-Martín, E Orduña Malea y JM Ayllon, pero el autor solicitado no aparece. Este caso en particular no se debe a una indización incorrecta, el autor buscado se encuentra en cuarto orden de publicación (ver Figura 3.11b) y el área destinada para los nombres de autores del registro de GS posee un límite el cual se superó.

Para el segundo caso, en la búsqueda por autor en *Google Scholar* para "Daniel Torres Salinas", un resultado es el que se exhibe en la Figura 3.12a, en este caso el nombre del autor buscado es mucho más específico que el que aparece en el listado de autores del registro, a saber, el nombre del autor buscado posee más términos que su correspondiente dentro del registro. Este caso no se debe a una mala indización, para esta publicación, los autores firmaron con nombres recortados y el motor los indizó de esa manera. Pero el registro mostrado en la Figura 3.12b si obedece una indización incorrecta, pues solo algunos de los términos del nombre del autor buscado aparecen como partes del nombre de dos autores distintos.



(a) El nombre de autor buscado no aparece



(b) El número de autores supera el límite definido del registro

Figura 3.11: Registro de *Google Scholar* donde el autor buscado no aparece

(a) Caso del nombre recortado

(b) Caso del nombre en partes

Figura 3.12: Resultados de *Google Scholar* con nombres mal formados

Ambos escenarios representan un problema para las tareas de recolección, pues no solo se trata de nombres de autores que incluyen iniciales

en sus nombres (un problema que es medianamente sencillo de solucionar) sino que también existen nombres mal formados e incluso casos en los que el autor buscado simplemente no aparece como autor de la publicación a simple vista. Son en estos casos en los que el procesamiento inicial indica que estos registros no son comprobables. Para solucionar este inconveniente es necesario un re-procesamiento de estos registros para identificar si corresponden o no al autor buscado.

El re-proceso mencionado, consta de cuatro alternativas, una más específica que la siguiente:

- Alternativa 1: Se parte de un listado de todos los registros que necesitan un segunda comprobación, por cada uno de ellos se obtiene el listado de autores de la publicación, y por cada autor se obtiene el listado de todas las posibles combinaciones del nombre entregado por la función *CombinacionesDeNombres(string t)* antes mencionada. Para cada combinación del nombre se realiza una consulta a las tablas *CoAuthors* y *VariantsCoAuthors* para buscar si dicha combinación es una combinación ya almacenada anteriormente por otro registro, si es el caso se obtiene el *AuthorId* a partir del coautor o variante de coautor hallada, para indicar que el trabajo analizado corresponde a dicho autor. Estos casos se dan cuando se almacena un coautor o variante poco específico (con pocos términos o nombres) y luego se intenta almacenar el mismo nombre pero con más términos, al realizar las búsquedas en la primera pasada no se encuentra una combinación válida debido a la falta de especificidad del primer nombre almacenado.
- Alternativa 2: Si la **Alternativa 1** no dio resultado positivo, a saber, no se encontró un coautor o variante de coautor, se procede a realizar una consulta a la AK API con el título de la publicación del siguiente modo (ver Algoritmo 3.5):

```
queryString["expr"] = "Ti='"+
    Clean(Title.ToLower()) + "'";
```

Algoritmos 3.5 Consulta a AK API por Título de publicación

Las palabras que componen el título son normalizadas ya que la API no permite caracteres especiales ni de puntuación, palabras acentuadas o que contengan la letra "ñ" y los términos que componen el texto se deben convertir a minúsculas. La consulta a AK

API permite indicar la cantidad de registros a devolver, en este caso solo se necesitan los primeros 10 registros y se examinan los autores de cada uno con el objetivo de saber si contienen al *Nombres + Apellidos* o *Apellidos + Nombres* del autor buscado. Si es así, se selecciona el autor más productivo y se le asigna el registro analizado. Se elige arbitrariamente el autor más productivo debido a que este registro no pertenece a ninguno, y se parte del supuesto que es muy probable que dicho autor sea realmente el autor de la publicación.

- Alternativa 3: Si la **Alternativa** anterior no fue satisfactoria, se realiza una nueva consulta a la AK API tomando el título de la publicación examinada, pasando previamente por un filtro doble, el primer filtro es igual al aplicado en la **Alternativa 2** (se normaliza el texto para eliminar signos de puntuación, caracteres especiales, caracteres acentuados, letra *ñ* y se convierte todo el texto a minúsculas, el segundo filtro viene dado por la eliminación de términos utilizando una lista de *stop_ words* empleada en (Hug and Brandle, 2017) y disponible en el portal *GitHub.com*[7], el listado de palabras vacías cuenta con al rededor de 1500 palabras en distintos idiomas (Inglés, Francés, Alemán, Italiano y Español). Una vez aplicados ambos filtros, se construye una consulta de tipo *AND*-anidado utilizando el atributo de entidad "*W*" de *Microsoft Academic* (W = *Words*, palabras del título/resumen de la publicación para la búsqueda de texto completo). Por ejemplo para el título:"Data for free: Using LMS activity logs to measure community in online courses", la consulta construida quedaría de la siguiente forma: "And(And(And(And(And(And(And(And(W='data', W='free'), W ='lms'), W ='activity'), W ='logs'), W ='measure'), W ='community'),W = 'online'), W ='courses')". Se realiza la consulta obteniendo los primeros 10 resultados y se examinan los autores de cada uno con el objetivo de saber si contienen al *Nombres + Apellidos* o *Apellidos + Nombres* del autor buscado. Si es así, se selecciona el autor más productivo y se le asigna el registro que está siendo analizado.

- Alternativa 4: Si ninguna de las alternativas anteriores produjo resultado favorable, lo último que resta por hacer es una compro-

[7] https://github.com/eprintsug/microsoft-academic

bación en la web mediante la *Bing Web Search API*[8] del buscador *Bing*. La utilización de esta API también requiere una suscripción para obtener una *key* al igual que la *Academic Knowledge API*, los límites y condiciones para la utilización de forma gratuita se resumen en la Figura 3.13.

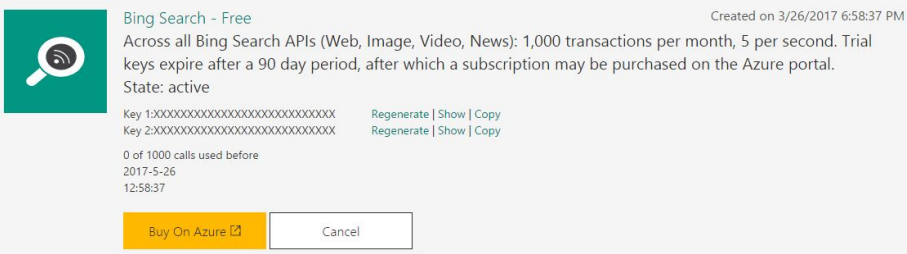


Figura 3.13: Restricciones para el uso de las claves de la *Bing Web Search API*

La consulta a la API se realiza con el título de la publicación normalizado como en la **Alternativa 2**, nuevamente solo se obtiene el top 10 de resultados, cada resultado es un conjunto de siete elementos (*id, name, url, about, displayUrl, snippet* y *dateLastCrawled*), de estos datos el valor que interesa para este trabajo es la URL al recurso que apunta, esto es así ya que las respuestas vienen dadas en un formato similar al presentado por la interfaz del buscador *Bing*. Cada URL es procesada mediante una consulta *HTTP Web Request* para recuperar el código HTML o el texto (en caso de ser un archivo PDF). El código o texto recuperado es utilizado para explorar dentro del mismo la presencia de los *Nombres + Apellidos* o *Apellidos + Nombres* del autor analizado, por medio de un conjunto de complejas expresiones regulares bien formadas para evaluar todas las posibles variantes de los nombres/apellidos del autor (caracteres acentuados, presencia de signos de puntuación, espacios en blanco, entre otras cosas). Si la búsqueda fue positiva, el registro se asigna al autor más productivo del conjunto de variantes halladas. Por el contrario, y luego de haber analizado las cuatro alternativas presentadas, este registro se marca como no comprobable definitivamente. La Figura 3.14 presenta un esquema

[8] `https://azure.microsoft.com/es-es/services/cognitive-services/bing-web-search-api/`

del proceso de consulta, recuperación de resultados y procesamiento de datos descripto.

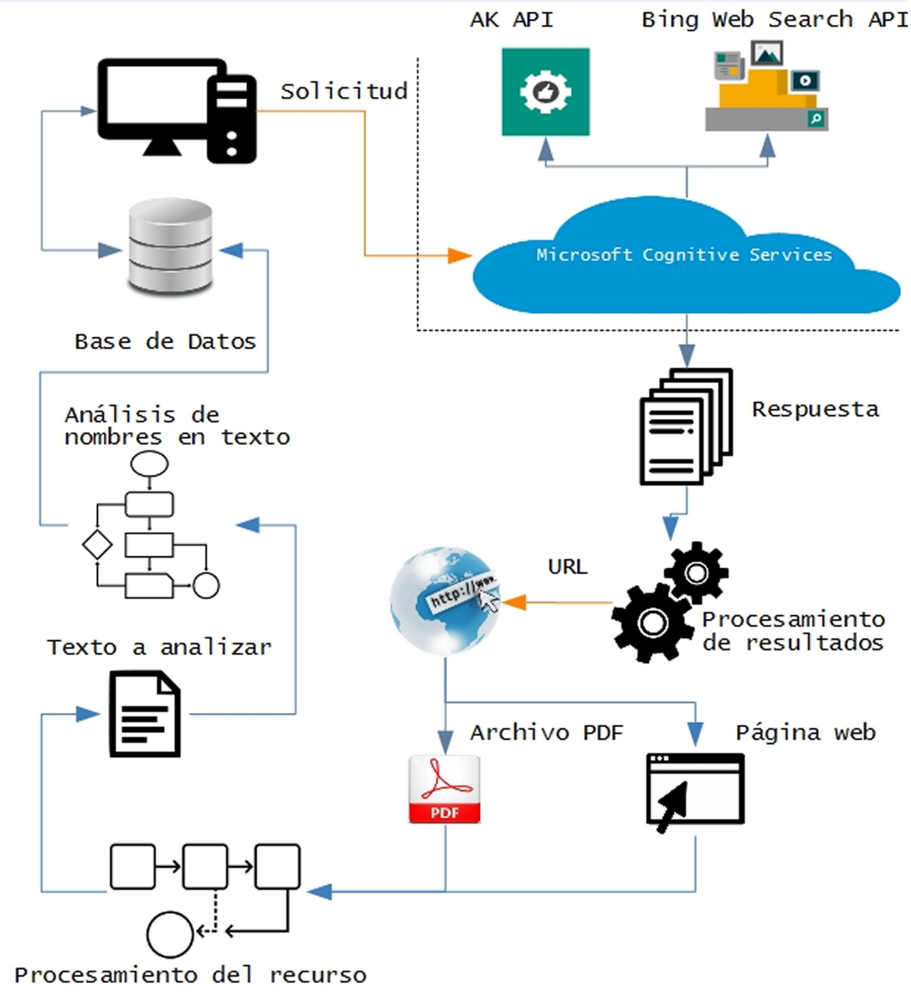


Figura 3.14: Esquema del proceso de consulta y recuperación de información mediante las APIs

Se mencionó que son los coautores los "responsables" de formar o determinar las distintas agrupaciones o conjuntos de coautores, estos grupos están dados por la relación que poseen con el autor analizado.

Salta a la vista que autores que escriban con personas distintas la mayor parte del tiempo, tendrán varias agrupaciones mientras que autores que suelen escribir con un grupo reducido o común de personas tendrán menos agrupaciones. Esto es así debido al esquema utilizado para agrupar los diferentes conjuntos de coautores en común que comparten publicaciones con el autor analizado. Existirán tantas agrupaciones como conjuntos de coautores "no relacionados", este supuesto puede no ser del todo cierto, ya que dos coautores que no comparten una publicación no es indicio que el autor no sea la misma persona, pero el uso de la coautoría para identificar estas agrupaciones puede arrojar dicha conclusión. Sin embargo se pueden emplear algunos esquemas, no para forzar la unión de dos o más conjuntos, sino para hallar alguna señal de que estos grupos si pueden estar relacionados. Por ejemplo, utilizando información de la web para ampliar la red de conocimiento del autor (Kang et al., 2009; Stefano et al., 2013; Abdulhayoglu and Thijs, 2017). Esquemas como el conducido por (Han et al., 2004, 2005; Gurney et al., 2012; Liu et al., 2014a,b, 2015; Lerchenmueller and Sorenson, 2016; Eltermann et al., 2016) utilizan información adicional como las palabras en los títulos de las publicaciones, palabras claves, nombres de revistas, lugar de publicación, información geográfica, encabezados de materias, lenguaje de publicación, datos filiatorios, direcciones de correo electrónico, datos de los resúmenes, fecha de publicación y características del nombre de autor.

En este trabajo, siendo que los datos analizados provienen de una fuente poco normalizada y poco controlada, y solo se cuenta con un conjunto finito de datos relativamente obligatorios, no es posible realizar la desambiguación como en los trabajos mencionados anteriormente, en otros términos, no es posible utilizar datos geográficos, ni filiatorios, ni del campo de estudio (salvo utilizando técnicas estadísticas y de procesamiento de lenguaje natural) para romper esa suerte de "empate" entre dos o más conjuntos. Al contar con pocos datos, la estrategia empleada fue determinar la existencia de algún patrón de colaboración entre dos grupos de coautores distintos y aparentemente desconocidos entre si, buscando entre ellos la existencia, ya sea de alguna publicación no incorporada en el conjunto de registros recolectados, es decir, siendo autores o coautores de una misma publicación (la manera más sencilla y directa de hallar una relación), o por el contrario analizando si algunos de estos coautores se conocen o algún coautor de los coautores es común a ambos grupos. La hipótesis planteada es la siguiente: si dos coautores de dos

grupos distintos (grupos **A** y **B**) poseen algún trabajo/publicación en común sin importar si incluyen o no al autor objeto de la búsqueda, se concluye que ambos conjuntos de publicaciones, las del grupo **A** y las del grupo **B** fueron escritas por el mismo autor, caso contrario se concluye que son dos personas distintas (caso del homónimo o grupos de coautores no relacionados), pues al no contar con datos adicionales para resolver la ambigüedad no es posible llegar a una conclusión distinta.

El esquema aquí diseñado es muy similar al empleado por (Kang et al., 2009) basado en la expansión de coautores. Existen dos estrategias bien diferenciadas *author-centric implicit coauthors* (aIC) y *coauthor-centric implicit coauthors* (cIC). En la primera de ellas (aIC) la idea es utilizar el nombre del autor analizado (autor analizado $\Rightarrow$ **a** = "Enrique Orduña Malea") como *pivot* e ir buscando relaciones entre los distintos coautores (listado de coautores del grupo $Group_1 = \{C_1, C_2, C_3, C_4....\}$ = {"A Martin Martin","D Torres Salinas","J Luis Ortega","J Serrano Cobos",}) tomando grupos de dos en dos ({a, C_1}, {a, C_2}, {a, C_3}, {a, C_4}. En la segunda estrategia (cIC), la idea es armar las combinaciones de dos en dos entre los integrantes de ambos grupos ($Group_1$ y $Group_2$), dicho de otra forma, hallar relaciones entre los coautores, lo que se conoce como coautores de coautores. Ya sea que se utilice un esquema u otro, aumentar el número de restricciones (en vez de tomar grupos de dos en dos, tomarlos de a tres, de a cuatro o más) puede reducir la posibilidad de encontrar resultados favorables. Algo que se percibe claramente es que el esquema aIC no es aplicable al dominio de este problema, ya que todos los coautores de todos los grupos encontrados poseen alguna relación con el autor objeto de estudio, ya que fueron autores o coautores del conjunto de publicaciones procesadas, en este sentido, el esquema utilizado en este trabajo es el de cIC.

El proceso inicia estableciendo las distintas combinaciones de grupos de coautores a evaluar tomados de dos en dos, estas combinaciones se arman de forma escalonada partiendo del listado de autores encontrados previamente, ordenados de forma descendente por cantidad de publicaciones. Luego por cada combinación ($Comb_1 = Group_1$ y $Group_2$) se obtienen dos conjuntos (set_1 y set_2), un conjunto por cada grupo. El set_1 está compuesto por el coautor más prolífico de $Group_1$ y el set_2 está compuesto por los dos coautores más prolíficos de $Group_2$ (si solo hubiese un coautor en este segundo grupo se tomará el único). Con ambos conjuntos, se toma el coautor de set_1 y el primer coautor de set_2 y

se utiliza la *Bing Web Search API* para recuperar los primeros 10 resultados con el parámetro de búsqueda: *"(Nombres y Apellidos de coautor* set_1*) AND (Nombres y Apellidos de coautor*$_1$ set_2*)"*.

Solo se recuperan páginas web, esto es, se filtran el resto de contenidos (imágenes, videos y noticias). Por cada resultado se obtiene la URL del recurso, y con dicha URL se realiza la recuperación del contenido de la misma (*web scraping*), en esta se revisa el código HTML o el texto en PDF para encontrar ambos coautores. Para lograr ello se utilizan expresiones regulares primero para descartar que los nombres de estos coautores se encuentren en secciones como el resumen/abstract, introducción, y principalmente las citas bibliográficas del recurso obtenido (bajo el supuesto, de que en la mayor parte de los resultados, al tratarse de investigadores, se recuperará material científico-académico). Una vez descartadas las secciones del contenido web que no se desea analizar, se revisa, mediante expresiones regulares, solamente si los apellidos de los dos coautores están presentes en el texto. Si esto es afirmativo, ambos conjuntos se fusionan, caso contrario, de existir un segundo coautor de set_2, se repite el mecanismo. Nuevamente, si el resultado es afirmativo, es decir, los apellidos de ambos coautores están presentes, estos conjuntos se fusionan, caso contrario, estos dos grupos se marcan como no satisfactorios y serán tratados en una revisión más exhaustiva. Este proceso se repite hasta analizar todas las combinaciones de grupos de coautores, cuando una combinación formada por dos grupos se fusiona, esta es candidata para una próxima evaluación con otro grupo distinto que haya sido satisfactoriamente evaluado. Los grupos marcados como no satisfactorios no participan en las revisiones sucesivas de la primera etapa.

La segunda etapa involucra solamente a los grupos marcados como no satisfactorios, para estos, por cada grupo ($Group_1$) se obtienen dos conjuntos (set_1 y set_2), el set_1 está compuesto por el coautor más prolífico de $Group_1$ y el set_2 está compuesto por hasta 15 de los coautores más prolíficos que no pertenecen a $Group_1$. Nuevamente se utiliza la API de *Bing* para recuperar los primeros 10 resultados con el parámetro de búsqueda: "*(Nombres y Apellidos de coautor* set_1*) AND (Nombres y Apellidos de Autor analizado)*". Por cada resultado obtenido el procedimiento es el mismo que el detallado anteriormente, solo difiere en que al momento de revisar el código HTML o el texto recuperado, se evalúa la presencia de los apellidos del coautor de set_1 contra los apellidos de al

menos uno de los coautores de set_2. A la primer coincidencia se determina que ambos conjuntos deben fusionarse, por el contrario, luego de recorrer todo el listado de set_2 y no encontrar coincidencias, se concluye que ambos grupos son distintos con lo cual o se trata de un homónimo o simplemente no existe ninguna relación entre ambos grupos de coautores.

El procedimiento anterior se generaliza en el Algoritmo 3.6

```
//primera recorrida
while (true)
{
  var listAuth = db.Select_AuthorsId(pattern, source);
  for (int k=0; k<listAuth.Count; k+= 2)
  {
    Combination comb = new
       Combination(listAuth[k].id, listAuth[k+1].id);
  }
  var lstComb = db.GetCombinations();
  foreach (Combination c in lstComb)
  {
    var set1 = db.Sel_CoAuthorId(1, c.Id1).ToList();
    var set2 = db.Sel_CoAuthorId(2, c.Id2).ToList();
    bool isFound = await CheckBingAPI(set1.Name+" AND
       "+set2.Name1);
    if (!isFound && set2.Name2!=null)
      isFound = await CheckBingAPI(set1.Name+" AND
         "+set2.Name2);
    if (isFound)
     MergeSets(c.Id1, c.Id2);
  }
  //si solo hay dos autores distintos, una única
     recorrida
  if (listAuth.Count == 2)
    bContinue = false;
}
//segunda recorrida
var lstNo = db.GetGroupsNo();
foreach (Group c in lstNo)
  {
    var set1 = db.Sel_CoAuthorId(1, c.Id1).ToList();
    var set2 = db.Sel_CoAuthorIdExcept(15,
       c.Id1).ToList();
    int iResult= await CheckBingAPI(set1.Name+" AND
       "+AuthorSearch,set2);
    if (isFound)
```

```
    MergeSets(iResult, c.Id1);
}
```

Algoritmos 3.6 Chequeo de relaciones entre coautores

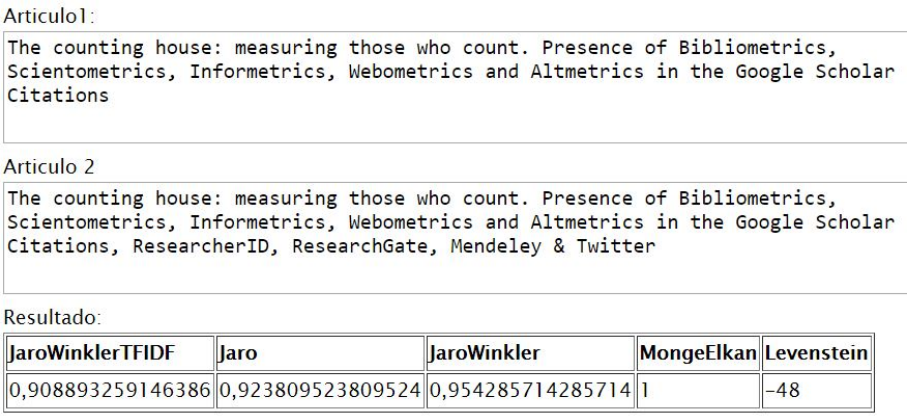
3.3.2. Detección de duplicados

El objetivo principal de cualquier esquema de detección de duplicados es establecer las reglas de validación por un lado y por otro lograr reducir los tiempos de análisis. Para este tipo de problemas, esquemas supervisados con conjunto de datos de entrenamiento no resultan adecuados, debido a la gran variabilidad que presentan las publicaciones de un autor a otro. En tal sentido es casi obligatorio emplear un esquema en tiempo real para la detección y posterior resolución del problema.

Anteriormente se había mencionado que las comprobaciones de todos contra todos (comprobaciones 1 a n) son prohibitivas debido a la gran cantidad de tiempo que consumen ($0(n^2)$ orden cuadrático) , en vista de ello los esquemas que entregan resultados más que aceptables resultan ser los que emplean algún tipo de agrupación, *blocking* o *clustering*. Para este trabajo se empleó una agrupación utilizando el año de la publicación, he aquí la importancia de contar con este dato y porque se descartaban los registros que inicialmente no poseían año de publicación. Bajo el supuesto de que una publicación duplicada no difiere en el año de publicación, se toma este esquema para dividir los conjuntos de publicaciones, por el contrario, si una publicación que posee todos los datos iguales pero difiere en el año de publicación, se considerarán como dos publicaciones distintas.

Una vez filtrados y descartados los registros con los que no se trabajará (ya sea que no corresponden al autor o el usuario decidió no incluir algunas de las variantes de nombres ofrecidas), estos se agrupan por año de publicación y sobre estos se realiza el análisis de duplicados utilizando la función de similitud *JaroWinklerTFIDF* (Cohen et al., 2013a), un esquema poco rígido basado en la métrica de distancia *TFIDF*, extendida para utilizar una técnica de *token-matching* con la métrica de distancia *JaroWinkler* (Winkler, 1999). Esta función de similitud difiere de *MongeElkan*, que fue utilizada para comparar dos nombres de autores, en que es más sensible a pequeños cambios o diferencias morfológicas, y además, al evaluar una cadena contenida en otra no arroja un valor de

coincidencia de 100 % como lo haría la función *MongeElkan* (ver Figura 3.15).


Articulo1:

The counting house: measuring those who count. Presence of Bibliometrics, Scientometrics, Informetrics, Webometrics and Altmetrics in the Google Scholar Citations

Articulo 2

The counting house: measuring those who count. Presence of Bibliometrics, Scientometrics, Informetrics, Webometrics and Altmetrics in the Google Scholar Citations, ResearcherID, ResearchGate, Mendeley & Twitter

Resultado:

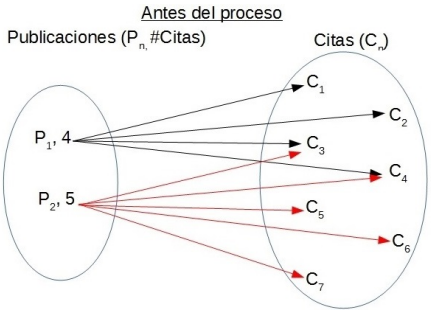
JaroWinklerTFIDF	Jaro	JaroWinkler	MongeElkan	Levenstein
0,908893259146386	0,923809523809524	0,954285714285714	1	-48

Figura 3.15: Valores de distintas funciones de edición al comparar dos artículos

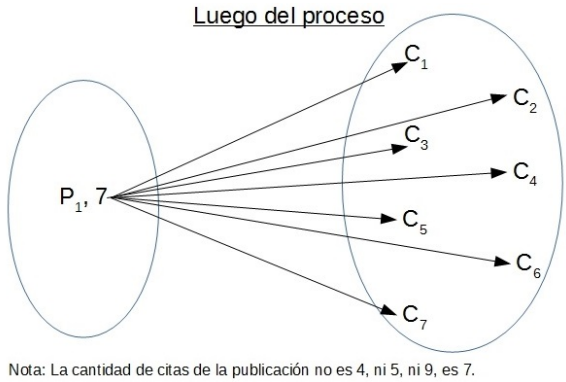
Luego de algunas pruebas se fijó el umbral de comparación en 0.92, si la comparación de dos artículos publicados el mismo año supera el valor de 0.92, son considerados duplicados. Para determinar con precisión la cantidad de citas se pueden seguir dos alternativas dependiendo de la información almacenada:

- Opción 1: Si los artículos poseen los registros de citas almacenados, lo que se hace es una fusión de ambos conjuntos y sobre estos se descartan los duplicados aplicando el mismo esquema de comparación arriba mencionado. De esto se obtendrá un número de citas más cercano a la realidad (observar Figura 3.16).

- Opción 2: Si no se cuenta con los registros de citas de estos artículos, se tomará el máximo número de citas registrado por cualesquiera de los registros duplicados. Se recuerda que además de duplicados pueden existir triplicados, cuadruplicados y más copias de un mismo registro, para estos casos el esquema sigue la misma lógica.



(a) Antes del proceso de fusión



(b) Resultado del proceso de fusión

Figura 3.16: Esquema de fusión de citas para publicaciones duplicadas

3.3.3. Comparaciones con diferentes valores de umbrales

Las distintas funciones de distancia de edición entregan un resultado/*score* como solución al comparar dos conjuntos de cadenas de caracteres. Como se enunció en el **Capítulo 2**, existen funciones propias y otras que son combinaciones de dos o más funciones, cada una de ellas tienen incorporada la lógica necesaria para evaluar con precisión y no todas son aplicables a todos los problemas, sino que depende de la naturaleza del mismo. Por ejemplo, para cadenas cortas y de pocos términos, con la posibilidad de que una de las cadenas evaluadas esté totalmente contenida dentro de la otra o al menos contengan parte de los términos de una en otra, la función *MongeElkan* entrega resultados muy favorables, por ello la elección de esta para comparar dos nombres de autores. En cambio, para comparar cadenas largas que pueden o no estar conte-

nidas una en otra, o que difieren algunos términos y esta variación si es sensible, la función que mejor se adapta es la de *JaroWinklerTFIDF*, que es la elegida para comparar dos títulos de publicaciones con el objeto de detectar posibles duplicados.

Pero la tarea no acaba con el solo hecho de elegir que función aplicar, sino que se deben fijar los umbrales de comparación, esto es, los valores que de ser superados al momento de comparar contra los resultados de aplicar la función de distancia de edición para dos cadenas de caracteres, las cadenas serán consideradas como duplicados. En la Tabla 3.3 se presentan algunos resultados luego de comparar títulos de publicaciones con varias funciones de distancia de edición.

Las funciones de distancia de edición entregan como resultado un valor que va de [0, 1], siendo 0 el valor que indica que dos cadenas no son para nada similares y 1 el valor que indica que dos cadenas son muy similares o idénticas. La única función que no sigue esta regla es la función de distancia de edición *Levenshtein* (Levenshtein, 1966), la cual entrega 0 si dos cadenas son exactamente iguales y un valor entero que indica la cantidad de caracteres a cambiar para convertir una cadena en otra.

Ya se había adelantado que el umbral de comparación para la función *JaroWinklerTFID* era de 0.92, pues como lo indica la Tabla 3.3, es el valor que mejor encaja a la hora de detectar posibles duplicados, esto no quita la eliminación de falsos negativos, pero se parte del supuesto de que esta cantidad solo representa un pequeño porcentaje. En este mismo sentido, el valor de umbral elegido tampoco libera al proceso de comparación de incorporar falsos positivos, nuevamente, se presupone que este número solo representa un pequeño porcentaje.

Para fijar los valores de los umbrales de la función *MongeElkan* para comparar dos nombres de autores, se siguió el mismo procedimiento indicado anteriormente, en la Tabla 3.4 se recogen un conjunto de pruebas de los valores de las funciones de distancia de edición para comparar variantes de un nombre de autor. Los valores utilizados en los procedimientos de evaluación se establecieron en:

- umbralNombre = 0.92
- umbralNombre2 = 0.80

Tabla: 3.3: Valores de funciones de distancia de edición para distintos artículos

Artículo 1	**Artículo 2**	**Jaro Winkler TFIDF**	**Jaro**	**Jaro Winkler**	**Monge Elkan**	**Levenshtein**
From dilemmatic struggle to legiti-mized indifference	From dilemmatic struggle to legitimized indifference: expatriates' host country language learning and its impact on the expatriate-HCE relationship	0,49	0,77	0,86	0,96	-100
Compilación de Proyectos de Investigacion de 1984-2002	Compilación de Proyectos de Investigación desde el año 2003 al 2012	0,51	0,85	0,91	0,78	-19
Academic home pages	Academic home pages: Reconstruction of the self	0,65	0,80	0,88	1	-28
Fault diagnosis for a MSF using a SDG and fuzzy logic	Fault diagnosis for a MSF using neural networks	0,67	0,82	0,89	0,68	-18
Scholarly use of the Web: What are the key inducers of links to journal Web sites?	Scholarly Use of the Web: What Are The Key...	0,78	0,81	0,88	0,93	-40
Índice H de las revistas científicas españolas según Google Scholar Metrics (2007-2011)	Índice H de las revistas científicas españolas según Google Scholar Metrics (2008-2012)	0,84	0,98	0,99	0,96	-2
Búsqueda y gestión de información científica en Ciencia y Tecnología	Búsqueda y gestión de información científica en Ciencias Sociales	0,88	0,93	0,96	0,87	-10
Rankings ISI de las universidades españolas según campos y disciplinas científicas (2011)	Rankings ISI de las universidades españolas según campos y disciplinas científicas:(2ª edición 2011)	0,89	0,95	0,97	0,93	-12

Artículo 1	**Artículo 2**	**Jaro Winkler TFIDF**	**Jaro**	**Jaro Winkler**	**Monge Elkan**	**Levenshtein**
Wikipedia as a tool for introducing social concerns into science education	Wikipedia as a tool for introducing social implications into science education	0,90	0,89	0,93	0,60	-9
Búsqueda y gestión de información científica en Ciencia y Tecnología. 2ª ed.	Búsqueda y gestión de información científica en Ciencias Sociales. 2ª ed.	0,91	0,92	0,95	0,86	-10
Spanish monolingual track: the impact of stemming on retrieval	Spanish monolingual track: the impact of stemming on retrieval.(2002)	**0,94**	0,96	0,97	1	-7
Diseño de un motor de recuperación de información para uso experimental y educativo	Diseño de un motor de recuperación de la información para uso experimental y educativo	**0,95**	0,91	0,94	0,57	-3
Edición electrónica de Informes de la Construcción y Materiales de construcción: datos prelimiares visibilidad y difusión.	Edición electrónica de" Informes de la Construcción" y" Materiales de Construcción": datos preliminares, visibilidad y difusión	**0,95**	0,81	0,88	0,29	-27
Las cifras de la enseanza universitaria en Documentacin en Espaa: 2006	Las cifras de la enseñanza universitaria en Documentación en España : 2006	**0,99**	0,97	0,98	0,51	-4
Categorizacin automtica de documentos en espaol: algunos resultados experimentales	Categorización automática de documentos en español: algunos resultados experimentales	**0,99**	0,98	0,99	0,49	-3

Artículo 1	**Artículo 2**	**Jaro Winkler TFIDF**	**Jaro**	**Jaro Winkler**	**Monge Elkan**	**Levenshtein**
Contenidos del buscador Google. Distribucin por pases, dominios e idiomas	Contenidos del buscador Google: distribución por países, dominios e idiomas	**0,99**	0,98	0,98	0,55	-3
El año de las ciencias sociales y humanas	El año de las cincias sociales y humanas	**0,99**	0,94	0,96	0,97	-1
Academia.edu: Social network or Academic Network?	Academia.edu: Social network or Academic Network?: Social Network or Academic Network?	**1**	0,85	0,91	1	-37
La nueva lista de investigadores altamente citados de Thomson Reuters y el Ranking de Shanghai: situación de España y mapa universitario	La nueva lista de investigadores altamente citados de Thomson Reuters y el Ranking Shanghai: situación de España y mapa universitario	**1**	0,94	0,96	0,98	-3

Tabla: 3.4: Valores de funciones de distancia de edición para distintas variantes de nombres

Nombre 1	Nombre 2	Jaro Winkler TFIDF	Jaro	Jaro Winkler	Monge Elkan	Levenshtein
enrique orduna malea	malea orduna enrique	1,00	0,70	0,70	0,46	-14
enrique orduna malea	eo malea	0,41	0,00	0,10	0,83	-12
enrique orduna malea	enrique o malea	0,67	0,85	0,91	0,88	-5
enrique orduna malea	e ordunamalea	0,37	0,00	0,10	0,92	-7
enrique orduna malea	e malea	0,41	0,46	0,52	0,94	-13
enrique orduna malea	o malea	0,41	0,40	0,40	0,94	-13
enrique orduna malea	enrique ordunamalea	0,78	0,98	0,99	0,95	-1
enrique orduna malea	malea	0,58	0,42	0,42	1,00	-15
enrique orduna malea	orduna	0,58	0,44	0,44	1,00	-14
enrique orduna malea	enrique orduna	0,82	0,90	0,94	1,00	-6
enrique orduna malea	enrique orduna m	0,67	0,93	0,96	1,00	-4

3.4. Esquema propuesto para cuantificar la productividad

Existen numerosas formas de medir o evaluar la productividad de un científico, así mismo la gran cantidad de indicadores bibliométricos favorece la aplicación de estos de acuerdo al ámbito en el que se aplique. En lo que respecta a esta tesis, se deben seleccionar ciertos indicadores que ofrezcan una visión global de cuán productivo resulta ser un autor. Como se comentó en apartados anteriores, la productividad puede ser entendida y evaluada de diferentes modos, no siempre el autor que posee un mayor número de publicaciones es más importante que otro, el solo hecho de contabilizar la cantidad de publicaciones no es suficiente. En este mismo sentido, comparar dos autores de campos o especialida-

des distintas tampoco resulta una buena aproximación, ya que existen campos del conocimiento donde es más frecuente publicar que otros, o ciertas disciplinas donde lo más común es escribir un libro, a sabiendas del tiempo que lleva escribir y publicar un libro o un nuevo descubrimiento, mientras que otros científicos publican una enorme cantidad de artículos por temporada. Por estas razones, los esquemas basados en el recuento de citas son los más populares hoy en día. Medir la producción tiene que ver con el hecho de contabilizar las unidades producidas, en cambio medir la productividad tiene que ver con medir estas unidades producidas por unidad de tiempo, ambos conceptos están muy relacionados y existen casos donde la línea que los separa se vuelve muy fina. En este trabajo se evaluarán y medirán ambas magnitudes, sin hacer distinción si se habla de un concepto u otro.

3.4.1. Selección de indicadores

Los indicadores a calcular y visualizar serán:

- Cantidad total de publicaciones
- Número total de citas
- h-index
- $i10$-index
- g-index
- hg-index
- Cantidad de publicaciones duplicadas
- Cantidad y listado de de coautores.

3.4.2. Comparación con motores actuales

Cada base de datos o herramienta de evaluación ofrece un conjunto de indicadores, algunos de ellos son comunes a la mayoría, tal es el caso del h-index o el número total de documentos publicados y citas recibidas, otros solo aparecen en un sola herramienta, caso del g-index ofrecido

únicamente por *Publish or Perish*. A continuación y solo a modo ilustrativo, se ofrecen capturas de pantalla de las principales bases de datos académicas donde se puede apreciar los indicadores que cada una de ellas ofrecen.

Google Scholar, si el autor buscado/analizado posee cargado el perfil en *Google Scholar Citations*, la herramienta muestra el número total de citas, *h*-index y *h10*-index, como se aprecia en la Figura 3.17

Figura 3.17: Indicadores ofrecidos por *Google Scholar*

Por su parte, *Microsoft Academic* no ofrece ningún tipo de indicadores salvo el número total de resultados encontrados (ver 3.18), sin embargo la *Preview*[9] de la versión 2 de *Microsoft Academic*, ofrece el número total de artículos y el número total de citas, como se aprecia en la Figura 3.19.

Scopus se limita a contabilizar el número total de documentos, número total de citas, número total de documentos que citan las publicaciones, *h*-index y cantidad de coautores (ver Figura 3.20).

La *WoS* presenta un poco más de detalles que los anteriores, por su parte contabiliza el número total de documentos, el número total de veces que fue citado, total de veces citado sin citas propias, el número de artículos en que se cita, artículos totales en que se cita sin citas propias, promedio de citas por elemento y *h*-index (ver Figura 3.21).

Por último *Publish or Perish*, entrega un detalle un tanto más minucioso, ofrece el rango de años de publicación, cantidad total de publica-

[9] `https://preview.academic.microsoft.com`

ciones, cantidad total de citas, citas por año, citas por publicación, citas por autor, publicaciones por autor, autores por publicación, h-index, g-index, *hI,norm* y *hI, annual* (ver Figura 3.22).

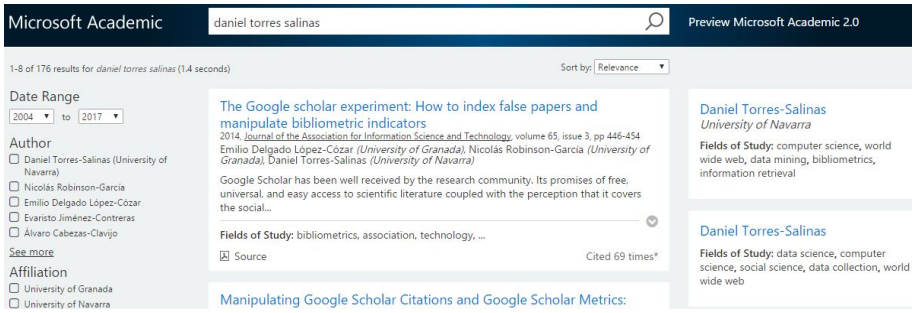


Figura 3.18: Indicadores ofrecidos por *Microsoft Academic*

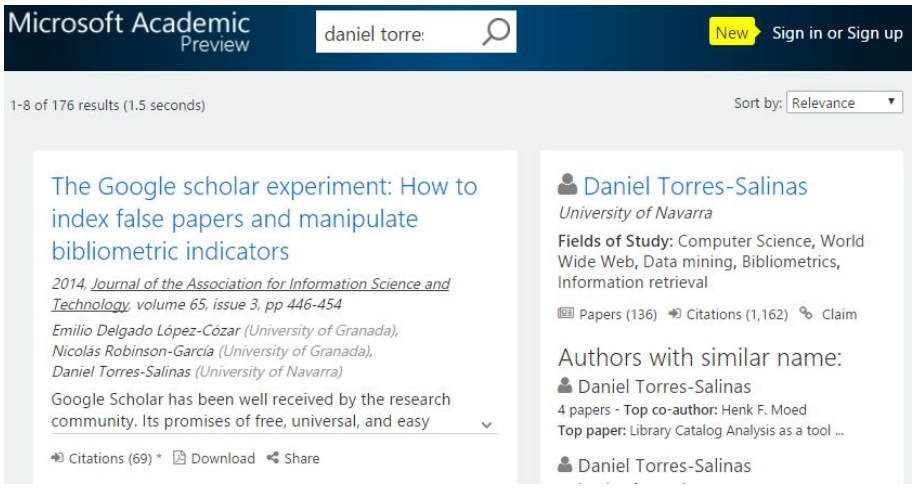


Figura 3.19: Indicadores ofrecidos por *Preview Microsoft Academic 2.0*

3.5. Técnica de Visualización Temporal

El diseño de visualizaciones para la exploración de datos temporales requiere varias opciones basadas en aspectos de tiempo y representación visual. La elección del diseño o técnica debe adecuarse al problema en

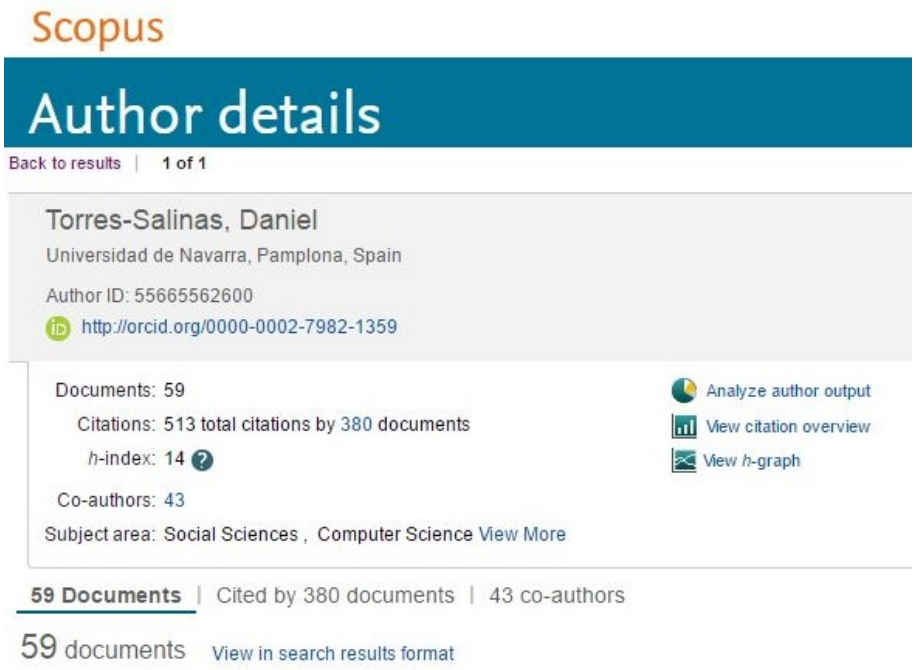


Figura 3.20: Indicadores ofrecidos por *Scopus*

cuestión y a los valores que se desean informar, la misma magnitud puede representarse de múltiples formas y el conocimiento que se pueda obtener de ella no tendrá el mismo impacto visual aplicando un diseño u otro. Como indica (Henkin and Dykes, 2016) los aspectos a tener en cuenta deberían basarse en el diseño, la forma y el tamaño de las marcas visuales.

La representación puede o no brindar alguna animación, como el cambio de posición de los elementos visualizados, sin embargo este no es un requerimiento ni una condición que deba robar mucha atención, pues solo agrega un efecto visual. Por otro lado y por ello no menos importante, un aspecto a tener en cuenta es el nivel de interacción que ofrezca la visualización, los elementos más utilizados son controles o barras deslizantes, listas desplegables o simples cajas de texto según las posibilidades que brinda el lenguaje de programación elegido, la interacción es muy importante pues las visualizaciones estáticas resultan poco atractivas sin un nivel de interacción, no se debe perder de vista que la interacción es un caso especial de animación donde el cambio está bajo el control del

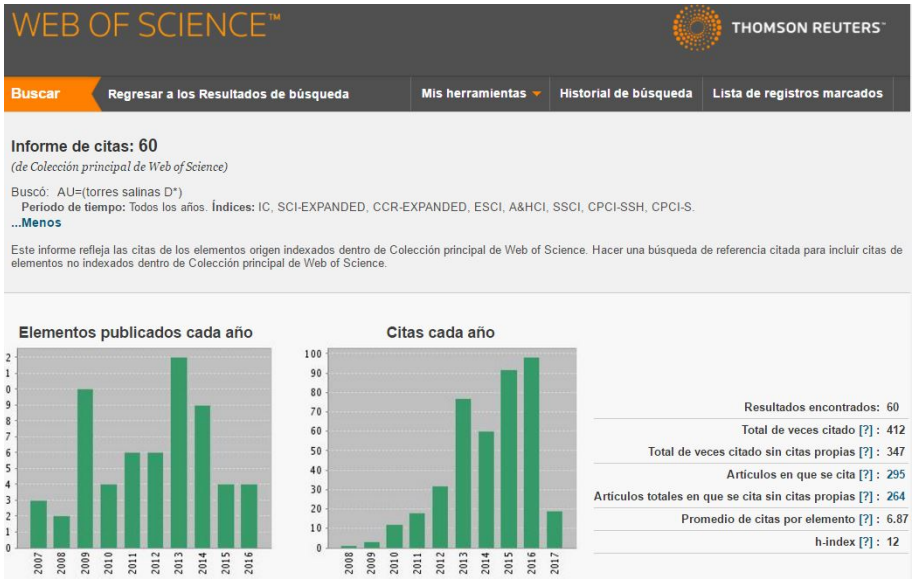


Figura 3.21: Indicadores ofrecidos por *Web of Science*

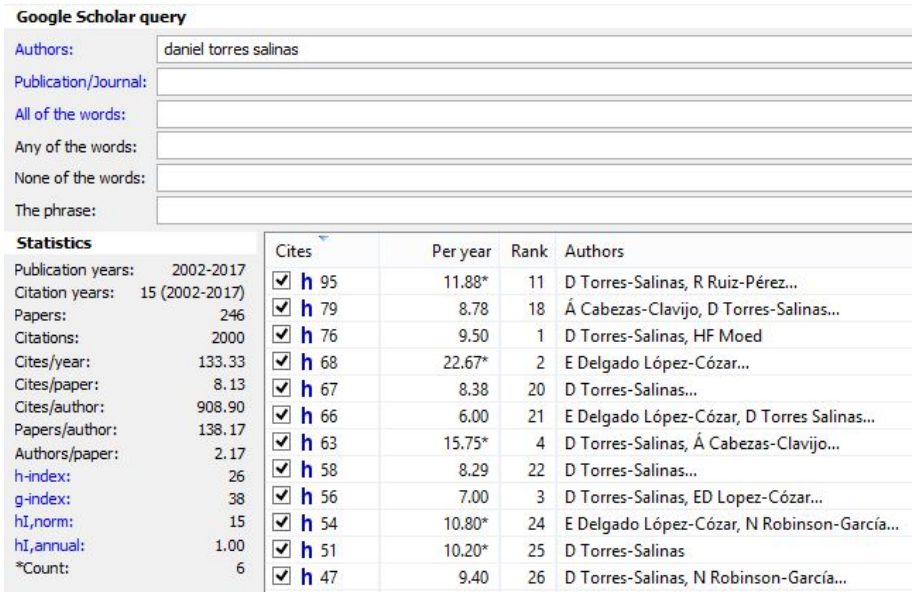
usuario.

En (Bach et al., 2014, 2016) se presenta una taxonomía completa de las distintas técnicas y modelos utilizados para representar datos temporales, tanto en 2 dimensiones (2D) como en 3 dimensiones (3D), dicho estudio se centra específicamente en las operaciones elementales que se pueden aplicar sobre el cubo espacio-tiempo, y explica cómo se pueden combinar para convertir un cubo espacio-tiempo tridimensional en una visualización bidimensional fácilmente legible.

3.5.1. Técnicas actuales

A continuación se presentan de forma breve algunas de las técnicas utilizadas para representar el tiempo como variable y magnitud, existen muchísimas otras y no es la idea detallar ni clasificarlas en este trabajo.

Quizás la forma más sencilla sea la lineal, indicando el paso del tiempo en uno de los ejes de un sistema cartesiano. La dificultad de este gráfico radica en su sencillez, solo se pueden representar dos magnitudes y quizás como se ve en la Figura 3.23, jugar con marcadores para indicar



Google Scholar query

Authors:	daniel torres salinas
Publication/Journal:	
All of the words:	
Any of the words:	
None of the words:	
The phrase:	

Statistics

Publication years:	2002-2017
Citation years:	15 (2002-2017)
Papers:	246
Citations:	2000
Cites/year:	133.33
Cites/paper:	8.13
Cites/author:	908.90
Papers/author:	138.17
Authors/paper:	2.17
h-index:	26
g-index:	38
hI,norm:	15
hI,annual:	1.00
*Count:	6

Cites	Per year	Rank	Authors
h 95	11.88*	11	D Torres-Salinas, R Ruiz-Pérez...
h 79	8.78	18	Á Cabezas-Clavijo, D Torres-Salinas...
h 76	9.50	1	D Torres-Salinas, HF Moed
h 68	22.67*	2	E Delgado López-Cózar...
h 67	8.38	20	D Torres-Salinas...
h 66	6.00	21	E Delgado López-Cózar, D Torres Salinas...
h 63	15.75*	4	D Torres-Salinas, Á Cabezas-Clavijo...
h 58	8.29	22	D Torres-Salinas...
h 56	7.00	3	D Torres-Salinas, ED Lopez-Cózar...
h 54	10.80*	24	E Delgado López-Cózar, N Robinson-García...
h 51	10.20*	25	D Torres-Salinas
h 47	9.40	26	D Torres-Salinas, N Robinson-García...

Figura 3.22: Indicadores ofrecidos por *Publish or Perish*

algún dato adicional. Sobrecargar el gráfico con múltiples curvas resulta engorroso, con lo que se debe cuidar que datos mostrar y cuales no.

Con los gráficos de barra (observar Figura 3.24) sucede algo muy parecido a lo que sucede con las representaciones lineales, aumentar la cantidad de tipos de barras para una nueva magnitud o para jugar con los colores es útil pero uno debe ser cuidadoso en no sobrecargar el gráfico, pues de lo contrario quedaría algo difícil de visualizar y complejo de interpretar.

Dejando un poco de lado las representaciones clásicas o comunes que todo el mundo conoce y utiliza a menudo, y entrando en lo que respecta al área de "Visualización de Información", la representación realizada por **Charles Joseph Minard** sobre la desastrosa campaña Rusa del ejército Napoleónico de 1812 (ver Figura 3.25), es según (Tufte, 1986), "el mejor gráfico estadístico jamás dibujado".

Este gráfico publicado en 1869 muestra diferentes variables en una única imagen bidimensional: la situación y dirección de las tropas, mostrando cómo las unidades se dividen y reagrupan; la merma de las tropas; el descenso de temperaturas y cómo éste influye en las bajas; la ubica-

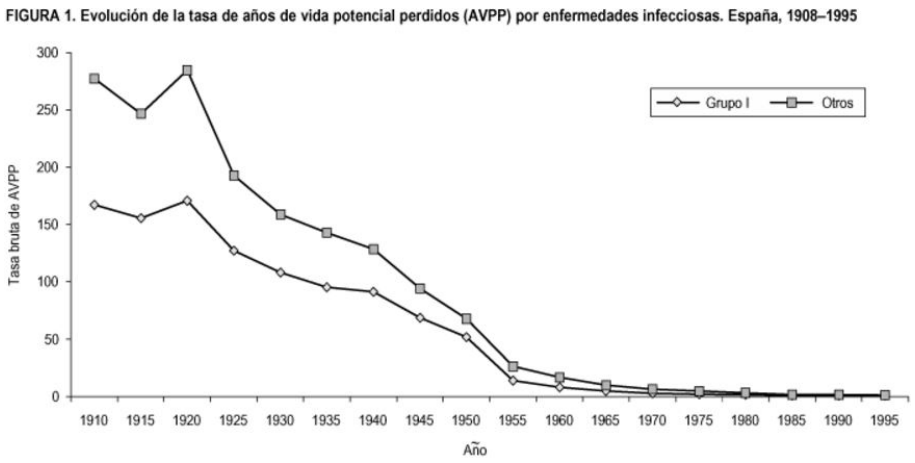


Figura 3.23: Gráfico de lineas tomado de (Canelo et al., 2002)

ción geográfica de los sucesos. En cuanto a representaciones de múltiples dimensiones es el mejor ejemplo además teniendo en cuenta que se desarrolló hace 148 años.

La representación realizada por Minard, es considerado el primer diagrama de *Sankey*, este tipo de diagrama es un tipo específico de diagrama de flujo en el que el ancho de los flujos representa la cantidad de flujo (Alemasoom et al., 2014) (ver Figura 3.26). Si bien guardan un parecido, el primer diagrama de Sankey fue publicado, oficialmente casi 30 años después de que Charles Joseph Minard expusiera su tan famosa representación, en el año 1898 por el capitán Matthew Henry Phineas Riall Sankey (Sankey, 1898; Schmidt, 2008).

La visualización de *Heat Map/Heatmap* (o mapa de calor según la traducción), es una representación gráfica donde los valores contenidos en una matriz son representados por colores, estos colores corresponden a una escala, que puede ser numérica o representar distintos estados, al ser una matriz posee dos dimensiones, una para las filas y otra para las columnas, en las cuales figuran los valores, escalas, o categorías que se desean visualizar. En la Figura 3.27 se puede ver un ejemplo de esta abstracción, donde las filas representan soluciones individuales, las columnas representan parámetros u objetivos y la escala de colores se normaliza para cada parámetro y objetivo.

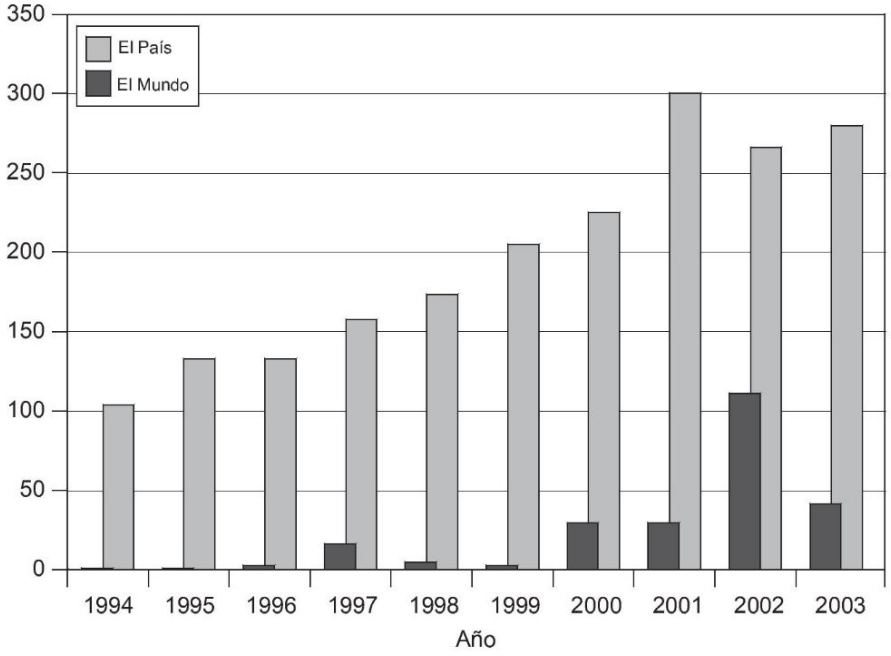


Figura 3.24: Gráfico de barras tomado de (Igual Camacho and Díaz Díaz, 2008)

Otro ejemplo de *Heatmap* es el que se observa en la Figura 3.28 utilizado para mostrar los patrones de migración de los países a lo largo de los años. El tiempo se organiza de forma lineal, con elementos de datos que se refieren a puntos instantáneos y se muestra una granularidad única. En un enfoque no jerárquico, el *heatmap* puede ser considerado como una disposición cartesiana en 2D compuesta por el tiempo y características de los objetos. El orden de ambos ejes no es obligatorio, pero si la magnitud lo requiere conviene hacerlo, como en el ejemplo mostrado, el eje temporal se ordena de forma ascendente del primer al último año.

La técnica *CircleView* es una combinación de técnicas de visualización jerárquica, tales como *Treemaps*, y técnicas de diseño circular, tales como gráficos de torta y segmentos de círculo (Keim et al., 2004). Como se aprecia en la Figura 3.29, la escala temporal inicia en el centro del círculo aumentando progresivamente hacia los márgenes del mismo. La disposición es polar con el ángulo asignado a un atributo ordinal, ordenado arbitrariamente, y la longitud del radio mapeado al año, en orden

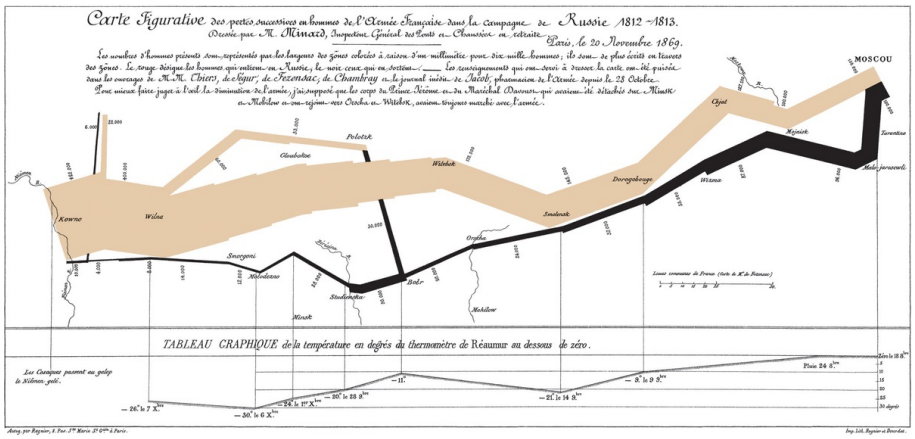


Figura 3.25: Gráfico de Minard de la campaña Rusa del ejercito de Napoleón 1812-1813, tomado de (Tufte, 1986)

ascendente ordinal. Esta configuración resulta en sectores ordenados hacia fuera desde el centro del círculo. El tamaño de los sectores se fija en relación con los datos, y el color también se asigna a una magnitud de un atributo (Henkin and Dykes, 2016)

El *Scatter plot* también llamado *scatter graph, scatter chart, scattergram, scatter diagram* o diagrama de dispersión, es un tipo de diagrama matemático que utiliza coordenadas cartesianas para graficar puntos que muestran la relación entre dos variables de un conjunto de datos. Los puntos pueden ser coloreados para indicar los valores de una variable adicional. Al ser puntos dentro de un eje de coordenadas, el valor de cada punto está dado por la posición que ocupa, es decir, el valor de una variable según la posición en el eje horizontal y el valor de la otra variable según la posición en el eje vertical.

A menudo se utiliza este tipo de diagramas para identificar asociaciones potenciales entre dos variables, en las que se puede considerar que una variable explicativa (como años de educación) y otra puede considerarse una variable de respuesta (como el ingreso anual) (Lacey, 2017). En la Figura 3.30 se observa un ejemplo de este diagrama donde en el eje vertical está indicado el crecimiento total del activo frente al crecimiento del aprovechamiento de estos en el eje horizontal.

Existen numerosos trabajos (Silva and Catarci, 2000; Aigner et al.,

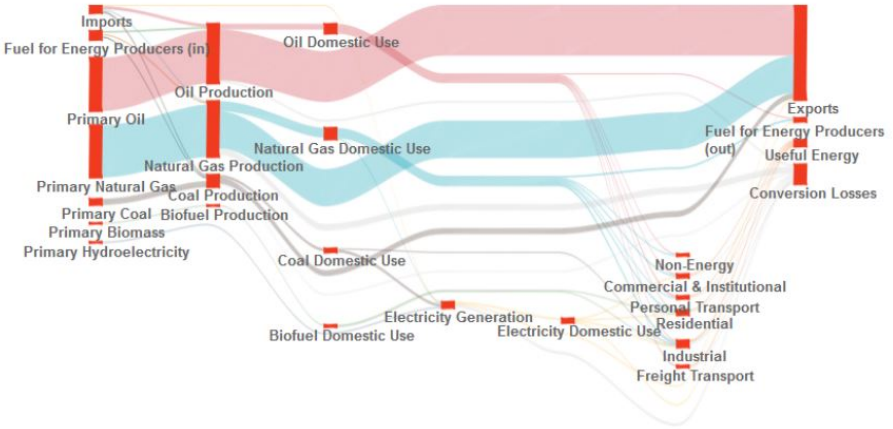


Figura 3.26: Ejemplo de diagrama de Sankey obtenido de (Alemasoom et al., 2014)

2007; Bach et al., 2014; Albo et al., 2016; Dornberger et al., 2016; Bach et al., 2016) relacionados con la representación del tiempo a través de diagramas, visualizaciones y herramientas, las técnicas aquí presentadas son solo un pequeño porcentaje de estas y no es la idea analizarlas ni discutir profundamente sobre las mismas. Si uno desea, podría utilizar un sin número de visualizaciones para representar el tiempo, porque de un modo u otro, el tiempo no deja de ser una magnitud medible y cuantificable, al menos por el momento.

3.5.2. Diseño de la visualización

Los conjuntos de datos que involucran más de dos magnitudes resultan difíciles de visualizar en un espacio de 2D, en algunos casos visualizaciones en 3D pueden aportar ciertas mejoras o nuevas posibilidades, sin embargo al aumentar la cantidad de magnitudes o dimensiones del problema resulta necesario recurrir a visualizaciones un tanto más complejas que una representación lineal en ejes de coordenadas. Para este trabajo se diseñó una visualización en el espacio 2D recurriendo a ciertos elementos para poder incluir todas las magnitudes que debían ser representadas para capturar, en una sola representación, el entendimiento global del estado actual de los resultados de la investigación de un

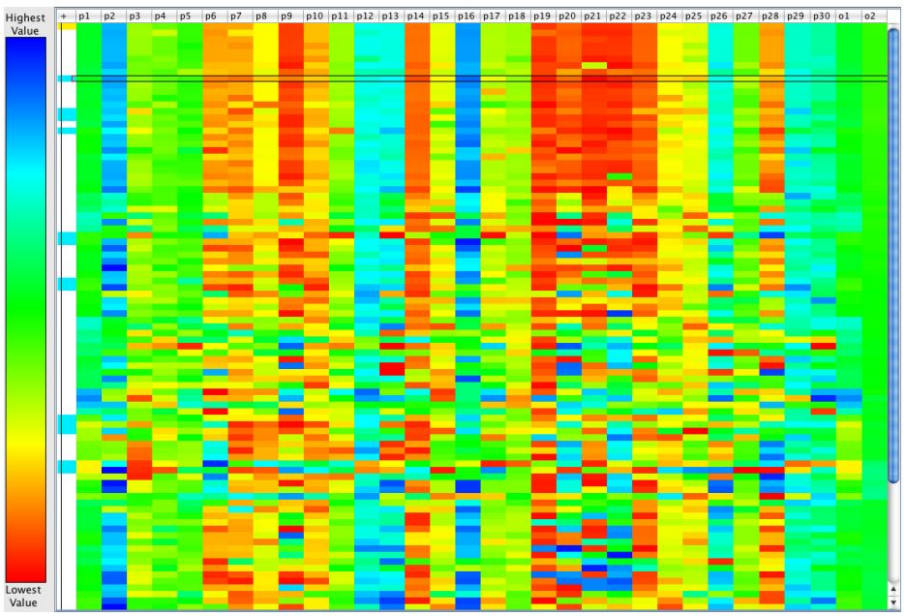


Figura 3.27: Ejemplo de HeatMap obtenido de (Hettenhausen et al., 2010)

científico.

Para dar cumplimiento a los objetivos planteados, en lo que respecta a la presentación de datos, se desarrolló una visualización utilizando la librería *D3.js* (Data Driven Documents) (Teller, 2013; Zhu, 2013) (ver Figura 3.31). Desde un principio, la idea de la visualización era que fuese interactiva y que abarcara la mayor cantidad de dimensiones para poder representar los datos de la mejor manera y no tener que ir transitando gráfico tras gráfico para tener un vistazo global.

Fundamentalmente, D3 es una elegante pieza de software que facilita la generación y manipulación de documentos web con datos. Lo realiza mediante (Murray, 2013):

- Carga de datos en la memoria del navegador.
- Vincular datos a elementos dentro del documento, creando nuevos elementos según sea necesario.

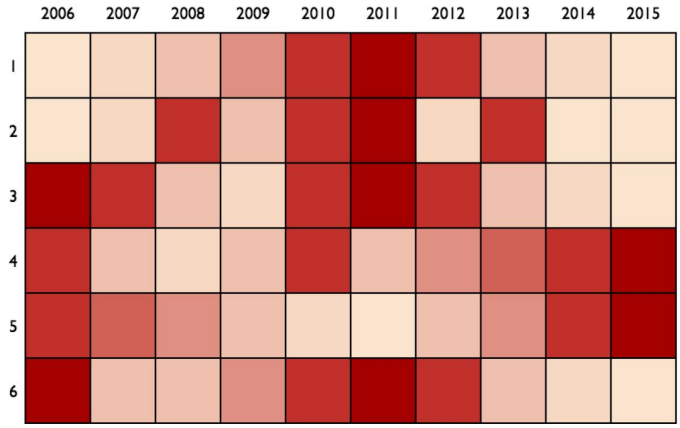


Figura 3.28: Ejemplo de HeatMap con escala temporal obtenido de (Henkin and Dykes, 2016)

- Transformar esos elementos mediante la interpretación de los datos enlazados de cada elemento y establecer sus propiedades visuales en consecuencia.
- Transición de elementos entre estados en respuesta a la entrada del usuario.

D3.js fue creado para llenar una necesidad apremiante de una sofisticada visualización de datos accesible desde la web. Debido al diseño robusto de la biblioteca, realiza más que simples gráficos. La *visualización de datos* ya no se refiere a los gráficos de torta y gráficos de líneas. Ahora significa mapas y diagramas interactivos y otras herramientas y contenidos integrados en noticias, cuadros de mando de datos, informes y todo lo que se ve en la web. El creador de D3.js, Mike Bostock[10], ayudó a desarrollar una biblioteca de visualización de datos anterior, *Protovis*[11], y también desarrolló *Polymaps*[12], una biblioteca de *JavaScript* que proporciona capacidad vectorial y composición de mosaicos en una forma ligera (Meeks, 2015).

Las posibilidades que entrega esta poderosa herramienta son inimaginables, si bien la curva de aprendizaje es empinada, la potencia y las

[10] https://bost.ocks.org/mike/
[11] http://mbostock.github.io/protovis/
[12] http://polymaps.org/

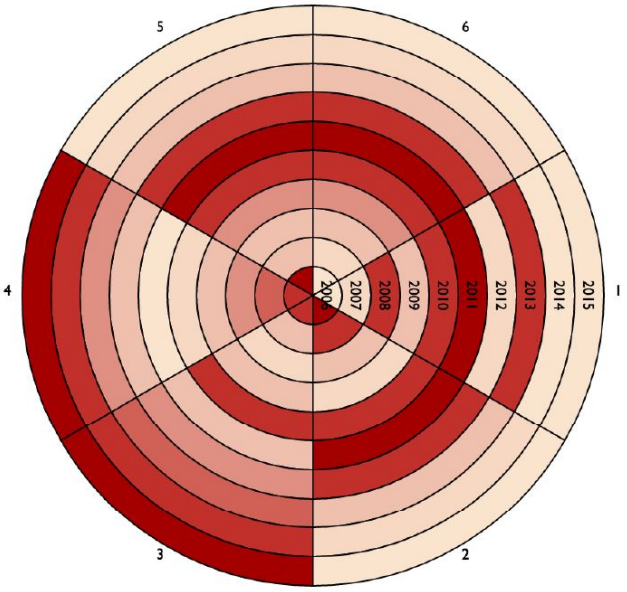


Figura 3.29: Ejemplo de *HeatMap* con escala temporal obtenido de (Henkin and Dykes, 2016)

capacidades de interacción que se pueden agregar en las distintas visualizaciones son enormes. La visualización desarrollada se basó en un ejemplo sencillo de *Scatter plot* disponible en el blog (Bostock, 2017). Este ejemplo se observa en la Figura 3.32 y carece de cualquier tipo de interacción, muestra solamente las disposición de los datos en el espacio cartesiano de 2D representando los valores de tres variables: longitud del sépalo, ancho del sépalo y tipo de flor de iris.

Luego de adaptar este ejemplo inicial, la visualización final es la que se observa en la Figura 3.33. Esta adaptación pone en juego varias magnitudes del problema, no solo con los elementos propios de este tipo de diagramas sino también a través de la interacción con la misma.

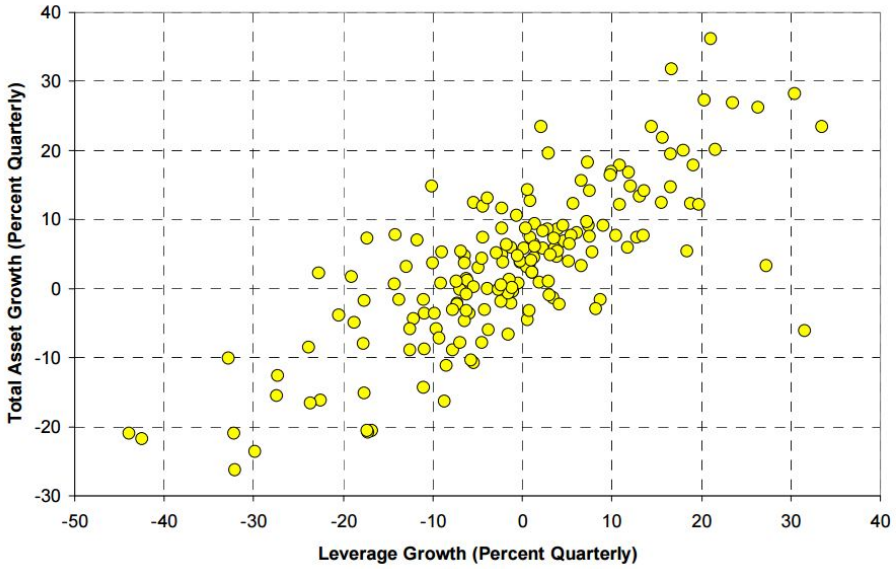


Figura 3.30: Ejemplo de *Scatter plot* obtenido de (Brunnermeier, 2009)

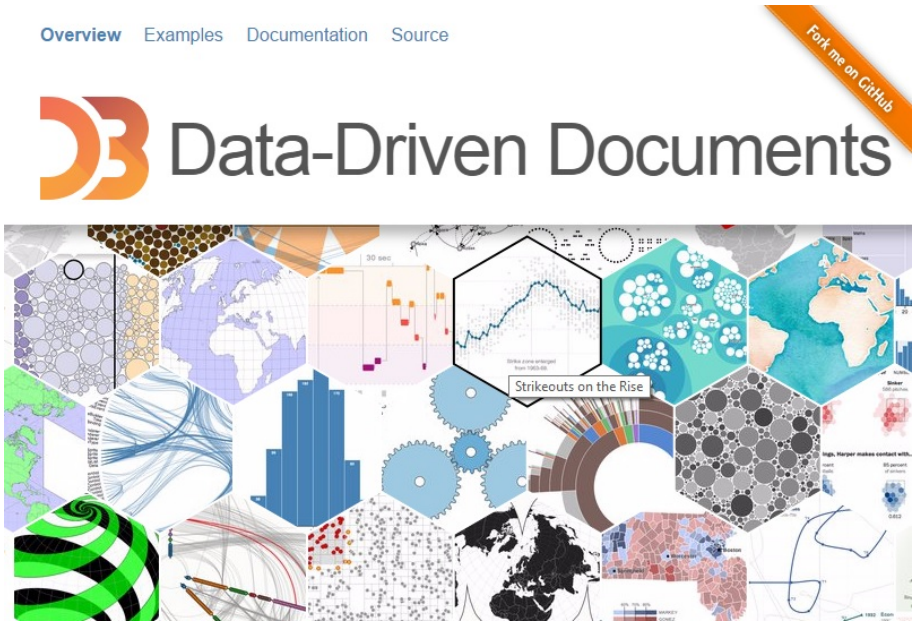


Figura 3.31: Portada del sitio web de D3

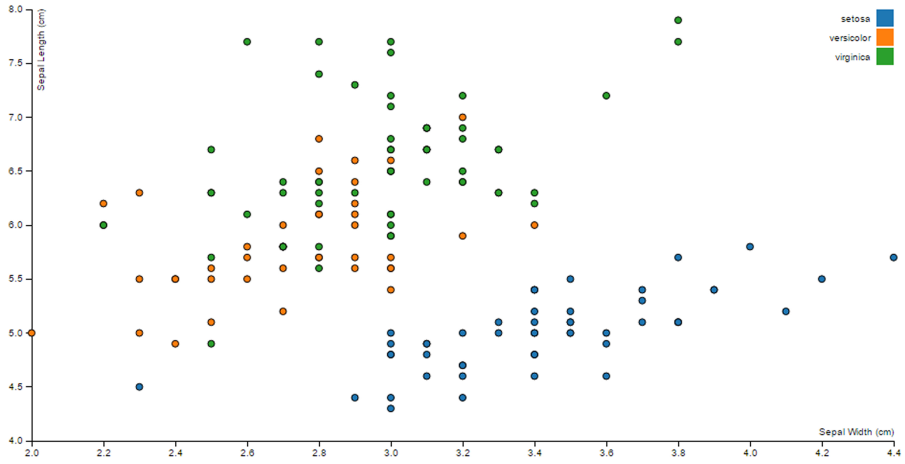


Figura 3.32: Visualización de *scatterplot* provistas por Mike Bostock

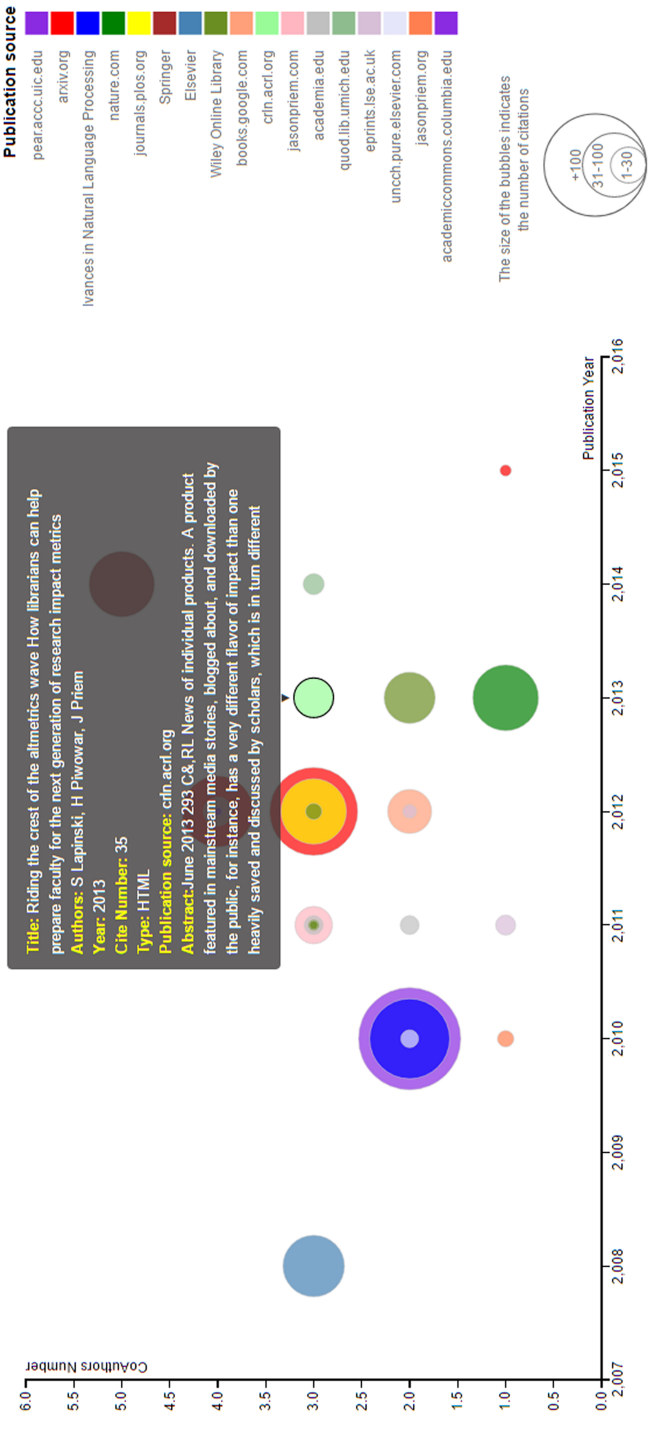


Figura 3.33: Visualización de *scatterplot* versión final

Luego de finalizado el proceso de desambiguación, queda un listado de registros de publicaciones "limpio" y listo para ser visualizado, la Figura 3.33 representa las publicaciones obtenidas de *Google Scholar* para Jason Priem [13]

En el eje vertical se encuentra la cantidad de coautores de las publicaciones, en el eje horizontal el año de publicación, el tamaño de las burbujas indica la cantidad de citas recibidas, y por último el color de las burbujas puede indicar varias cosas de acuerdo a los datos importados o recolectados: se pensó para que mostrase el nombre de la revista o *Journal* donde fue publicado, pero *Google Scholar* no ofrece este dato en todos los registros, por ello es que al realizar la recuperación para este motor, el campo almacenado es el *Publication source* (por ello el nombre del filtro visualizado), en caso de que los datos fueran recolectados de otras fuentes que si entregue el nombre del *Journal* (por ejemplo *Microsoft Academic* si entrega este campo), este campo se visualizará en el margen derecho de la pantalla con un color distinto para cada uno de ellos. Al ser un campo libre, el nombre del *Journal*, al realizar la importación de datos uno podría utilizarlo para categorizar cualquier elemento: por ejemplo el lugar geográfico de publicación, el tipo de evento donde fue publicado el trabajo (congreso, revista, capítulo de libro, jornada, conferencia), entre muchas otras posibilidades. No es un valor numérico, con lo cual solo acepta cadenas de caracteres como valores, sin embargo uno podría determinar una escala o rango de valores para una magnitud numérica y establecer los valores en cada registro a importar.

D3 posee un conjunto de opciones para crear e implementar animaciones e interacción con la visualización diseñada, algunas son sencillas de aplicar y otras requieren un esfuerzo considerable, además, estás se pueden combinar para proporcionar un efecto mucho más atractivo. Las animaciones están básicamente relacionadas con el cambio de posición de los elementos presentados o los efectos visuales para mostrarlos u ocultarlos. La interacción por su parte es muy similar a la animación, solo que los cambios en el contexto visual son operaciones que están bajo el control del usuario, es decir, es el usuario el que inicia la interacción con la herramienta mediante una acción, como hacer *clic* en un botón, como arrastrar el *mouse*, como hacer *scroll*, entre tantas otras. Para un mayor detalle de las opciones de animación e interacción otorgadas por D3 referirse a (Teller, 2013; Zhu, 2013; Murray, 2013; Meeks, 2015).

[13] `https://scholar.google.com/citations?user=w32jC0YAAAAJ`

La visualización ofrece seis posibles interacciones. La primera de ellas está relacionada con el elemento antes mencionado, el *Publication source*, este elemento se representa por un recuadro con un color distinto para cada ítem (ver Figura 3.34).

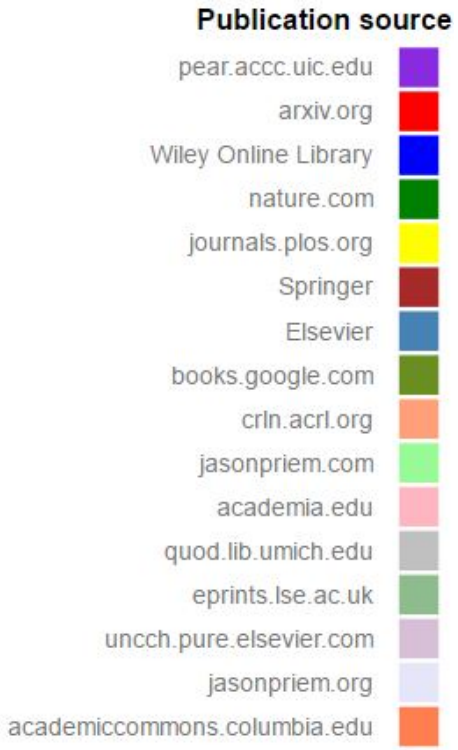
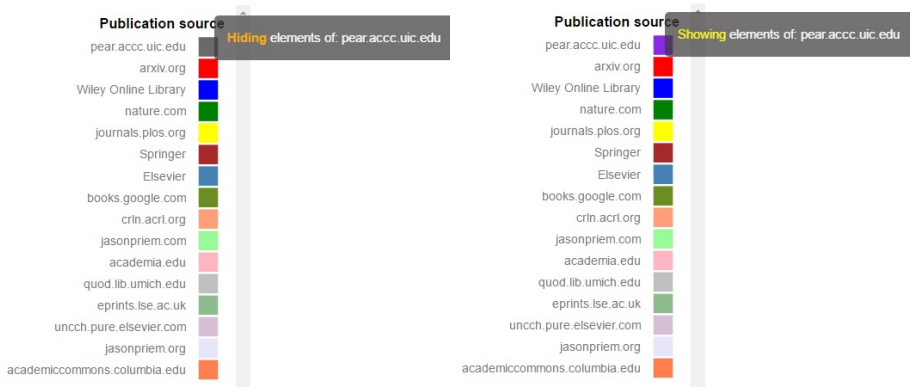


Figura 3.34: Filtro *publication source* de la visualización implementada

Al hacer clic sobre el recuadro de color de un elemento, la herramienta oculta las burbujas que poseen dicho valor o categoría (ver Figura 3.35a), una vez hecho esto la categoría se torna de color gris indicando que los elementos correspondientes están ocultos.

Al hacer clic nuevamente sobre el recuadro de color gris, la herramienta muestra las burbujas que poseen dicho valor o categoría (ver Figura 3.35b), una vez hecho esto la categoría retoma su color original indicando que los elementos correspondientes están visibles.

La segunda interacción que ofrece tiene que ver con el filtrado del



(a) Ocultar una categoría (b) Mostrar una categoría

Figura 3.35: Opciones del filtro *publication source*

formato de la publicación, es decir, con el tipo de archivo del recurso visualizado, la opción es un *OptionButton* que puede tomar cualquier de los siguiente valores: ALL, HTML, TXT, PDF, DOC, BOOK, PPT, XLS, PS u OTHERS, en la categoría OTHRES entran aquellos registros que no tienen indicado el formato o tipo de archivo, o aquellos que no han podido ser identificados como uno de los formatos mencionados. Como se observa en la Figura 3.36, la opción por defecto que muestra todos los elementos sin filtrar es "ALL", al hacer clic en cualquiera de los otros valores, las burbujas desaparecen o aparecen según se elija un tipo de archivo u otro.

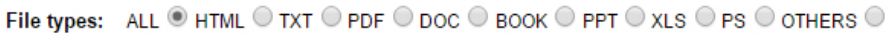


Figura 3.36: Filtro *File types* de la visualización implementada

Otra interacción es la posibilidad de filtrar los registros por año de publicación, a tal efecto el filtro es un control de tipo *Slider* como se aprecia en la Figura 3.37. Posee dos controles *slider*, uno para indicar el inicio del intervalo *From year* y otro para indicar el final del intervalo *To year*. Ambos filtros funcionan de forma independiente y las burbujas nuevamente, aparecen o desaparecen según posean un valor de año de publicación que esté dentro del rango de los filtros de año elegido.

La cuarta interacción es la posibilidad de visualizar el detalle de la

Figura 3.37: Filtro *From-To year* de la visualización implementada

publicación al pasar el mouse por encima de una burbuja, como se ve en la Figura 3.38 el detalle muestra: el título de la publicación (Title), listado de autores (Authors), año de publicación (Year), número de citas (Cite Number), tipo de archivo (Type), lugar de publicación/nombre de la revista/categoría (Publication source) y Resumen (Abstract). Al quitar el mouse del radio de la burbuja el detalle mostrado desaparece.

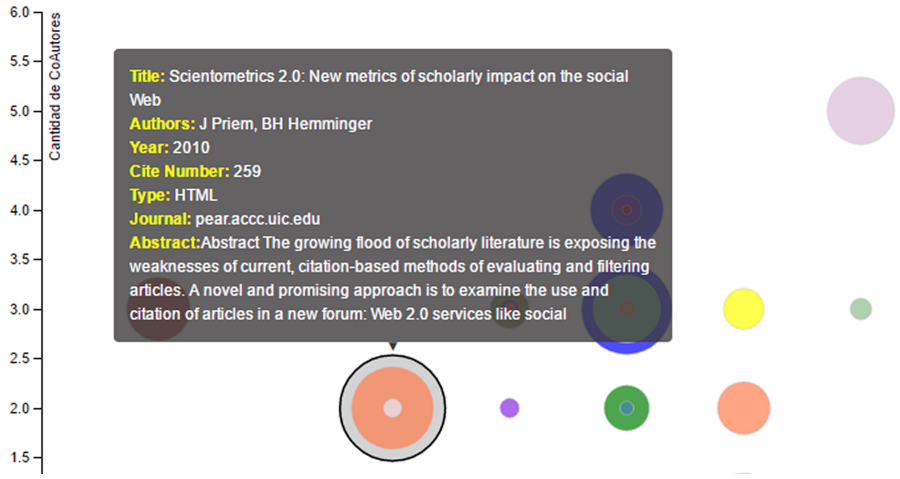


Figura 3.38: Detalle de la publicación al pasar el mouse sobre la burbuja

La quinta interacción permite filtrar las burbujas de acuerdo a la cantidad de citas recibidas, así pues existen dos controles de tipo *Slider* como se observa en la Figura 3.39. Un control para indicar el inicio del intervalo *From #Cites* y otro para indicar el final del intervalo *To #Cites*. Ambos filtros funcionan de forma independiente y las burbujas aparecen o desaparecen según posean un valor de cantidad de citas recibidas que esté dentro del rango los filtros de cantidad de citas elegido.

La última interacción es la posibilidad de acceder al recurso publicado por medio del link que se almacenó al momento de hacer la recuperación

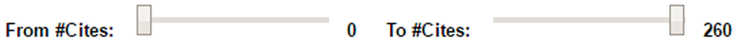


Figura 3.39: Filtro *From-To # Cites* de la visualización implementada

o importación de datos, para ello es necesario hacer doble click en cualquier parte de la burbuja, esto abrirá una nueva pestaña con el recurso solicitado.

La superposición de las burbujas se da cuando más de una publicación comparten el mismo año de publicación y la misma cantidad de coautores, sin embargo esta superposición es tal que las publicaciones con mayor número de citas se ubican al fondo mientras que las de menor cantidad de citas se ubican por encima.

Capítulo 4

Prototipo experimental *Academic-Evaluator*

Se describe en este capítulo el diseño del prototipo experimental que se desarrolló para validar y dar soporte al esquema de procesamiento y visualización de información propuesto y detallado en el **Capítulo 3**.

4.1. Diseño del prototipo

El prototipo de entorno visual llamado *Academic Evaluator-AE*, tiene como objetivo brindar el soporte necesario a todos los procesos involucrados en el tratamiento de la información que propone esta tesis. Si bien, la construcción de un prototipo no era uno de los objetivos planteados al iniciar este proyecto, la cantidad de procesos involucrados y la necesidad de tener todo centralizado en un mismo lugar, dio como resultado este desarrollo. El diseño modular del prototipo permitirá ampliar sus funcionalidades en etapas posteriores.

4.1.1. Arquitectura del modelo

El prototipo AE se ha diseñado como una aplicación web por dos motivos principales, primero por la gran cantidad de herramientas y *frameworks* disponibles para el desarrollo web, y segundo proporcionar acceso

a esta herramienta, a la comunidad científico-académica como al público en general, .

Los componentes principales de *Academic Evaluator* son:

- Procesos y lógica de negocio: el corazón del prototipo lo componen un conjunto de elementos y de librerías que logran procesar la información de forma transparente al usuario, desde una interfaz sencilla e intuitiva.
    - Motor lógico: sin lugar a dudas este es el componente principal, es quien posee toda la lógica de negocio y las reglas de inferencia para poder desambiguar un conjunto de publicaciones de manera automática. Posee un gran número de funciones y reglas para cubrir gran parte de los problemas planteados. Utiliza la potencia de las expresiones regulares para detectar patrones y la gran variabilidad tanto en los nombres de autores como en los títulos de las publicaciones.
    - *Secondstring*[1]: es una librería escrita en lenguaje *Java* que ofrece el compendio de funciones de distancia de edición más grande en la actualidad. De aplicación sencilla, se incorpora al prototipo desarrollado mediante la utilización de la librería IKVM[2] que permite la ejecución de librerías .JAR (extensión *Java*) dentro de un proyecto C#.
    - AK API: la *Academic Knowledge API* es la API de *Microsoft* que permite consultar y recuperar las publicaciones científicas almacenadas en *Microsoft Academic Graph*
    - *Bing Web Search API*: esta API permite realizar consultas utilizando el motor de búsqueda de propósito general *Bing*.
    - *Json.NET*: es un *framework* JSON para .NET, utilizado principalmente para serializar/deserializar objetos en formato JSON.
- Interfaz gráfica web diseño adaptativo: la interfaz gráfica se desarrolló adaptando el *template AdminLTE* de *Almsaeed Studio* [3] el cual posee un diseño adaptativo ideal para desarrollos web.

[1] `http://secondstring.sourceforge.net/`
[2] `https://www.ikvm.net/`
[3] `https://almsaeedstudio.com/`

- Visualización de información: la visualización principal se desarrolló bajo el *framework* de visualización D3, la misma fue creada únicamente para los propósitos de este trabajo.
- Procedimientos almacenados y consultas a la BD: se utilizó la potencia que brinda el motor de base de datos *SQL Server 2012 SP2* no solo para persistir los datos sino también para crear procedimientos almacenados y consultas con el objeto de mejorar la extracción de información a partir de los datos recolectados.
- Se utilizaron los lenguajes de programación: C#, JavaScript, HTML5, CCS3 y T-SQL.

La arquitectura del software desarrollado se presenta en la Figura 4.1.

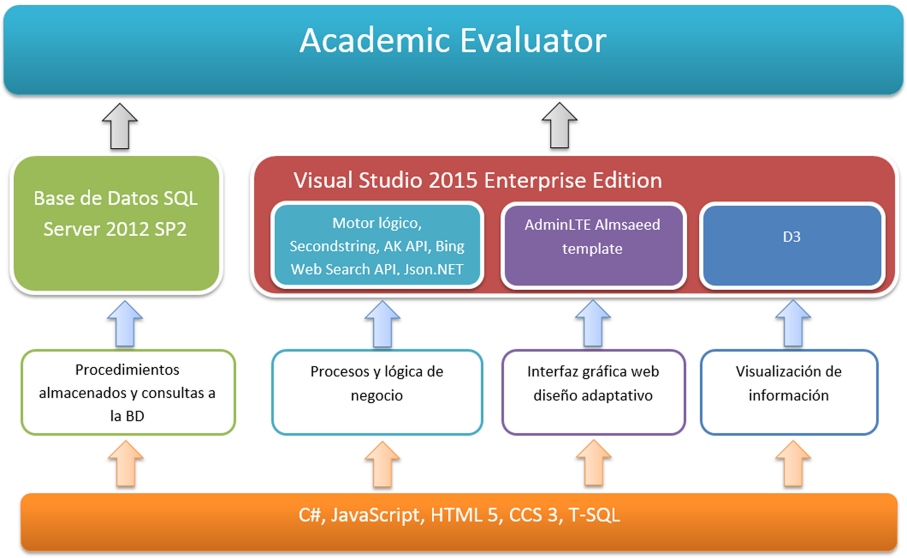


Figura 4.1: Arquitectura del prototipo *Academic Evaluator*

4.1.2. Diseño de interfaces

El prototipo web fue codificado en el lenguaje *C# 6.0* bajo el *Framework .NET v4.6*, utilizando como entorno de desarrollo *Visual Studio*

2015 Enterprise Edition. Es una aplicación web donde los elementos de la interfaz gráfica son una adaptación de la plantilla *AdminLTE*, la cual posee un diseño adaptativo que utiliza todas las propiedades y bondades ofrecidas por los lenguajes HTML5 y CCS3. La interacción y los efectos en la transición de una opción a otra o al momento de obtener resultados y respuestas por parte del sistema, resultan agradables y adecuados, y con un alto rendimiento tratándose de un aplicación web con un alto contenido visual.

Como se puede observar en la Figura 4.2 la pantalla inicial del prototipo consta de 4 secciones:

- Opciones de búsqueda: la única opción de búsqueda es por autor, para ello se deben ingresar los Nombres completos sin iniciales, seguido de los Apellidos completos sin iniciales y por último el origen de datos. El botón con el icono de la lupa inicia el proceso de búsqueda de registros dentro de la base de datos según los criterios ingresados.
- Panel de resultados inicial: abajo de las opciones de búsqueda, se ofrece un panel de resultados inicial, aquí se verán las alertas entregadas por el sistema y como se ve en la imagen, se listarán las agrupaciones de coautores encontradas.
- Acceso permanente a la búsqueda: este panel provee un acceso permanente a la búsqueda principal, se deben ingresar al igual que en el panel de búsqueda principal, los Nombres, Apellidos y origen de datos, separados por el caracter ";" (punto y coma).
- Panel de opciones: en el margen izquierdo se ofrece el acceso a las siguientes opciones:
  - *Data*: desde aquí se puede acceder a la opción de importar un nuevo archivo de publicaciones (opción *Import File*) y a la opción de revisar los distintos conjuntos de datos almacenados pertenecientes a los autores de los que se recolectaron o importaron los archivos de publicaciones (opción *Work With*).
  - *Charts*: una vez finalizada la desambiguación de registros (opción "Run disambiguation of records"), desde aquí se puede acceder a las visualizaciones del conjunto inicial de registros (opción *Before*) y del conjunto resultante de registros (opción *After*).

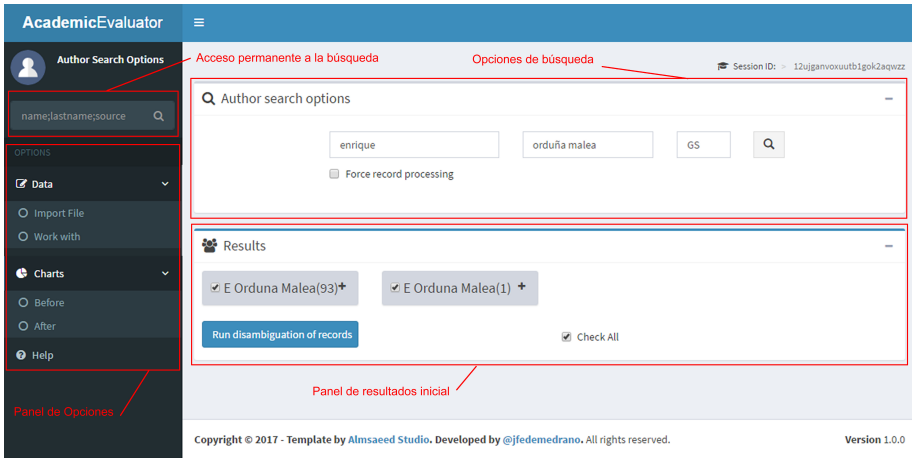


Figura 4.2: Pantalla inicial de *Academic Evaluator*

En la Figura 4.3 se presenta el panel principal de resultados, a este panel se accede una vez finalizada la desambiguación de registros. El mismo consta de 5 secciones:

- Nombre de Autor y variantes: se indica el nombre del autor analizado y las distintas variantes morfológicas del nombre encontradas en el procesamiento de las publicaciones.

- Gráficos estadísticos: se presentan dos gráficas conceptualmente sencillas, la primera de ellas ofrece un resumen de la cantidad de citas y de la cantidad de documentos publicados por año, desde el primer al último año de publicación almacenado en los registros del autor (gráfico de lineas, la curva gris representa la cantidad de citas recibidas y la curva azul representa la cantidad de documentos publicados), y la segunda gráfica ofrece un resumen de los formatos de archivo correspondientes a las publicaciones evaluadas en un gráfico de tipo *donut*, cada formato posee un color distinto y al pasar el mouse por la gráfica se puede observar el valor de cada área, que representa la cantidad presente de ese formato.

- Panel de indicadores: resumen de los indicadores ofrecidos por la aplicación, estos indicadores son: *Documents* (cantidad de publicaciones obtenidas luego del proceso de desambiguación), *Cites* (cantidad de citas recibidas por todas las publicaciones del conjunto

final), *Duplicates* (cantidad de registros duplicados encontrados), h-*index*, g-*index*, hg-*index* e i10-*index*.

- Acceso a las visualizaciones principales: al igual que las opciones del menú ubicado en el margen izquierdo de la pantalla, aquí también se ofrece acceso a las visualizaciones del conjunto de publicaciones tanto antes (opción *View chart before*) como después (opción *View chart after*) del proceso de desambiguación.

- Listados resultantes Documentos/Coautores: Por último se ofrece acceso al listado de publicaciones desambiguadas (*Documents*), sin incluir los elementos detectados como duplicados y aplicando la fusión de citas en caso de haber sido posible realizar tal procedimiento. La interfaz permite ordenar el listado por cualquiera de la columnas mostradas, también permite seleccionar la cantidad de registros a mostrar como también filtrar por cualquiera de las columnas y paginar los resultados. Por su parte el listado de Coautores (*CoAuthors*) muestra el conjunto de coautores resultante de procesar el conjunto de publicaciones final, se muestra el nombre del coautor y la cantidad de publicaciones en coautoría con el autor analizado, al hacer clic en el nombre del coautor se abrirá una ventana con el listado de dichas publicaciones.

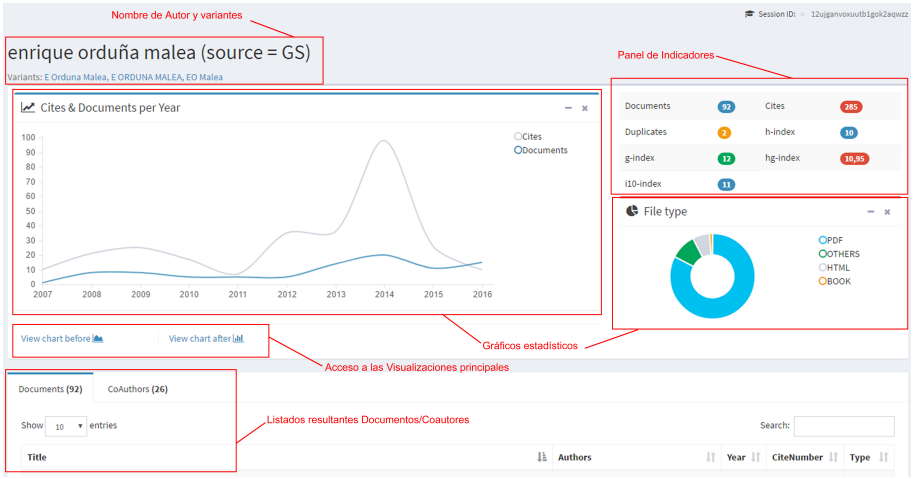


Figura 4.3: Panel de resultados de *Academic Evaluator*

4.1.3. Base de datos

Tanto los conjuntos de datos iniciales (importados o recolectados) como los resultados de los procesamientos llevados a cabo, son almacenados en una Base de Datos relacional que posee 11 tablas. La Figura 4.4 presenta un diagrama de la Base de Datos donde se puede apreciar el detalle de los campos de cada tabla y las relaciones que existen entre ellas. Las bases de datos relacionales poseen grandes cualidades no solo

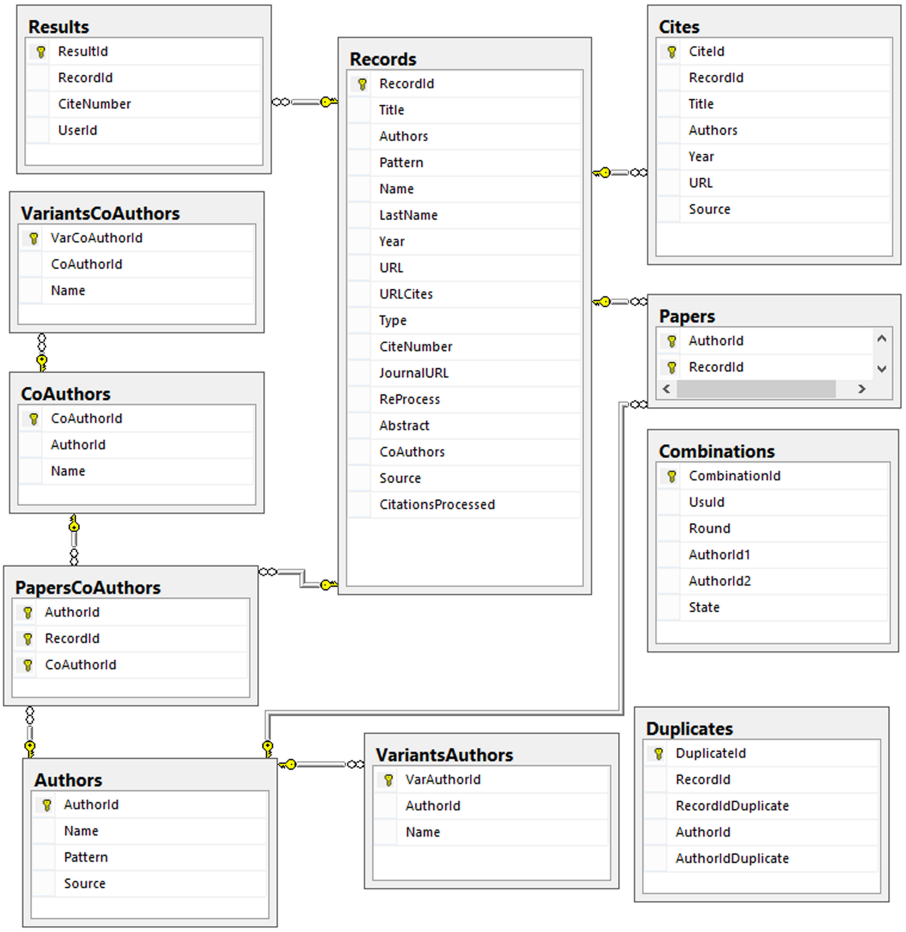


Figura 4.4: Diagrama de la Base de Datos provisto por el SSMS

para almacenar datos, sino también para establecer relaciones entre las

tablas, además, el motor de base de datos *SQL Server* tiene incorporadas muchísimas funciones que facilitaron en gran medida el procesamiento de los conjunto de datos. Se crearon procedimientos almacenados y funciones escalares para un gran número de procesos relacionados con el tratamiento de la información.

4.1.4. Visualizaciones

Contar con una visualización que resuma y/o aporte mayores conocimientos a los que se pueden ofrecer con tablas y gráficos estadísticos, resulta una alternativa viable. La idea de contar con visualizaciones es poder plasmar es un único gráfico, grandes conjuntos de datos o conjuntos de datos con muchísima información. Las herramientas que permiten construir visualizaciones de información, como el *framework* D3 que se utilizó, están pensadas para esta tarea.

Los efectos de transiciones y la interacción con el usuario son fundamentales, pues de otro modo serían gráficos estáticos que no aportan datos nuevos más allá de los que se pueden representar. Ya se habló puntualmente de todas las posibilidades que ofrece la visualización desarrollada en cuanto a niveles de interacción, y este es justamente el punto fuerte de la misma, permitir al usuario final jugar, probar, filtrar, acomodar, ocultar, mostrar no solo resulta un mecanismo atractivo, sino también reduce el trabajo que de otra forma debería de hacerse manualmente.

En algunos casos, el espacio de 2D puede limitar la representación de un gran número de variables o de un gran número de datos, por esta razón, las visualizaciones deben ser pensadas desde varias perspectivas y quien las diseña debe ser capaz de utilizar factores externos a la representación misma de modo de no sobrecargar el gráfico con datos abundantes e innecesarios. Las visualizaciones en 3D aportan una dimensión extra, pero muchas veces resultan complejas de diseñar y entender.

Si bien la visualización que se presenta en este trabajo no aporta muchos detalles en sí misma, sí ofrece un panorama general y gráfico del estado actual de la producción científica de un autor, que sumado al conjunto de indicadores ofrecidos y al detalle de los resultados, aportan un panorama completo y adecuado de gran parte del espectro de variables involucradas, y es el usuario quien decide qué ver y cómo verlo.

La aplicación plantea la misma visualización, con todos los elementos y opciones, en dos ámbitos diferentes.

4.1.4.1. Visualización anterior al análisis

Una vez finalizado el proceso de desambiguación, se crean los conjuntos de datos iniciales y resultantes para ser visualizados. La visualización antes del análisis (*Chart before process*) ofrece el panorama del conjunto inicial de datos, es decir, sin controles y sin filtros de duplicados, el usuario recibe una "fotografía" inicial, en esta opción no hay detalles ni indicadores sobre estos datos, pues estos se ofrecen sobre el conjunto definitivo de datos, sin embargo es una opción adicional simplemente para ver el punto de partida y poder compararlo con el resultado final.

4.1.4.2. Visualización posterior al análisis

En esta segunda visualización (*Graph after process*) comparada con la visualización anterior al análisis, pueden faltar elementos gráficos, desde burbujas (publicaciones) hasta las agrupaciones del elemento *Publication source*, esto es así puesto que los procesos de filtrado descartan y fusionan publicaciones de manera de ofrecer un conjunto de datos limpio y conciso, y es sobre estos datos que se calculan los distintos indicadores.

La razón de tener la misma visualización con dos conjuntos de datos relativamente "distintos" (no siempre serán distintos, pues si no existen datos a eliminar, ambos conjuntos serán idénticos) es poder ver la imagen del antes y el después, ver cómo cambian los datos y la representación de los mismos, el cambio será mayor cuanto menos normalizados estén los datos del conjunto de publicaciones, obviamente, de estar completamente normalizados los datos, y no existen registros duplicados, y no existen variantes morfológicas del nombre de autor y las agrupaciones pueden ser resueltas con las reglas de inferencias planteadas por el modelo, ambas visualizaciones serán exactamente iguales.

4.1.5. Opciones adicionales

El prototipo desarrollado ofrece en el panel de Opciones (*Options*), acceso a: *Import File* y *Work with*. La opción *Import File* permite impor-

tar un archivo en formato JSON con la estructura definida en el **Capítulo 3**. Este archivo contendrá el conjunto de publicaciones de un autor proveniente de cualquier origen de datos, además, si el origen de datos permite la recolección de las publicaciones que citan los documentos a incorporar, el archivo debe crearse de forma que refleje este conjunto adicional, que será como un *array* de elementos *Cites*. No existe un límite para el archivo a importar ni de cantidad de registros ni de tamaño. Como se observa en la Figura 4.5, además de indicar la ubicación del archivo a importar es necesario especificar los nombres (campo de texto *Author Name*) y apellidos (campo de texto *Author LastName*) del autor propietario del conjunto de publicaciones. Ambos datos (nombres y apellidos) junto al nombre del origen de datos, serán utilizados para recuperar los datos al momento de ejecutar los procedimientos correspondientes. El origen de los datos (campo *source*) viene incluido en el mismo archivo.

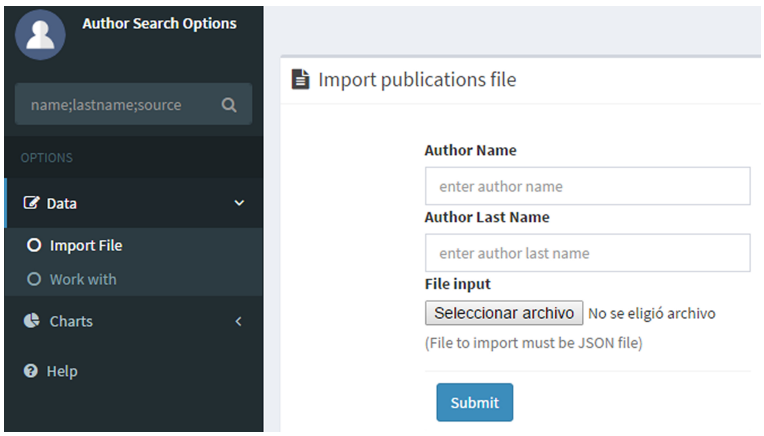


Figura 4.5: Opción *Import File*

Por otro lado, la opción *Work with* muestra en una tabla, los conjuntos de datos ya importados o recolectados por la herramienta, de este modo se realiza un recorrido por los elementos ya existentes y se obtiene así un rápido acceso para empezar a trabajar con ellos. La tabla posee opciones para ordenar los datos por cualquiera de las columnas y de filtrar los registros por cualquier valor mostrado. Así mismo posee opciones de paginación para no mostrar el conjunto completo de datos y sobrecargar la pantalla. Al hacer clic sobre el nombre del autor en cualquiera de las filas, la aplicación redirige a la pantalla inicial de búsqueda con

los criterios ya seleccionados.

List of patterns collected

Show 25 entries Search:

Select	Source	# Records
alvaro cabezas clavijo	GS	41
alvaro cabezas clavijo	MA	47
anne wil harzing	GS	119
anne wil harzing	MA	124
daniel torres salinas	GS	218
daniel torres salinas	MA	146
emilio delgado lopez cozar	GS	270
emilio delgado lopez cozar	MA	186
enrique e tarifa	GS	60
enrique e tarifa	MA	52
enrique orduna malea	GS	94
enrique orduña malea	MA	130
isidro f aguillo	GS	145
isidro f aguillo	MA	142
jason priem	GS	26
jason priem	MA	24
jason priem	MA	3
jose luis alonso berrocal	GS	160
jose luis alonso berrocal	MA	128
juan manuel ayllon	GS	30
juan manuel ayllon	MA	17

Figura 4.6: Opción *Work with*

Capítulo 5

Experimentos y Resultados

Para poder evaluar la herramienta desarrollada y los procesos diseñados, fue necesario llevar a cabo un experimento de principio a fin para ver cual es la realidad que ofrece el trabajo aquí realizado. Se realizará en dos partes, primero con datos provenientes de *Microsoft Academic* y luego con los datos obtenidos de *Google Scholar*. Se indicará el procesamiento realizado por cada conjunto de datos y posteriormente se señalará como se visualizan los resultados para ambas opciones. El experimento inicia con la recolección de datos de los orígenes mencionados (GS y MA), el procesamiento en cada una de sus etapas, la presentación de los resultados y la posterior evaluación de los mismos. Cada proceso mencionado en este capítulo fue oportunamente descrito en el **Capítulo 3**, por esta razón no se darán tantos detalles del funcionamiento en este capítulo, pero sí se exhibirán los resultados entregados por cada uno.

5.1. Experimentos con *Microsoft Academic*

La herramienta construida, *Academic Evalutor* (AE), brinda soporte al modelo de tratamiento de datos que se planteó y diseñó en esta tesis, la misma puede funcionar con cualquier origen de datos, el único requisito es que se respete el formato de importación de datos definido con anterioridad. Se describen a continuación la serie de pasos para recuperar, procesar y visualizar los datos, desde el conjunto de publicaciones de un autor a recuperar desde *Microsoft Academic*.

El autor elegido para el experimento es Enrique Orduña Malea[1]. Es Ingeniero Técnico de Telecomunicaciones, Licenciado en Documentación, Master en Contenidos Multicanal y Doctor en Documentación por la Universidad Politécnica de Valencia (UPV). Actualmente trabaja como Investigador postdoctoral en el Instituto de Diseño y Fabricación (IDF) de la UPV, y como profesor externo en el Departamento de Comunicación audiovisual, Documentación e Historia del Arte (DCADHA) de la misma universidad. Desde 2012 es miembro del Grupo de investigación EC3 (Evaluación de la Ciencia y Comunicación Científica) de la Universidad de Granada (UGR). Por otra parte, desde 2008 es miembro del Grupo ThinkEPI (y redactor del Anuario ThinkEPI), y desde 2011, es representante del COBDCV en el CT50/SC1 de Aenor (Agencia Española de Normalización y Certificación). Sus líneas de investigación se centran fundamentalmente en la Cibermetría, tanto descriptiva (testeo de indicadores de naturaleza web y unidades de análisis) como instrumental (análisis de fuentes y buscadores) y aplicada (principalmente a entornos de creación y consumo de información científica)[2].

5.1.1. Recuperación de información

Para iniciar el experimento es necesario contar con un conjunto de datos iniciales correspondientes a las publicaciones de un autor, *Microsoft Academic* es un motor académico que ofrece una interfaz de búsqueda de libre acceso pero no ofrece la posibilidad de exportar los resultados de la búsqueda. Tampoco permite recuperar los datos realizando *scraping*, para lograr este objetivo, MA ofrece la *Academic Knowledge API (AK API)* para recuperar el listado completo de publicaciones según los criterios de búsqueda que se le indiquen, para este caso serán las búsquedas por autor. *Academic Evaluator* hace uso de esta API y recupera los datos del autor objeto de análisis, el acceso a la AK API requiere de un *token* de autenticación, el cual fue previamente obtenido. *Microsoft* ofrece acceso gratuito a esa API por el término de un mes y luego requiere el pago por uso a través de la plataforma *Azure*[3], para los experimentos del prototipo descrito se utilizó la suscripción gratuita.

[1] https://scholar.google.com/citations?user=g6bEUdkAAAAJ

[2] http://www.cervantes.es/bibliotecas_documentacion_espanol/para_bibliotecarios/jornadas/jornada_6/cv_orduna_enrique.htm

[3] https://azure.microsoft.com/es-es/

La Figura 5.1 muestra la opción inicial de la herramienta AE, al no encontrarse resultados previamente almacenados con esos criterios de búsqueda, el sistema informa con una alerta y solicita que se intente la búsqueda con la opción de forzar el procesamiento de registros. Para este caso inicial, la opción de "procesamiento" no produce ningún resultado pues no existen registros almacenados, dicha opción es útil cuando se importaron nuevos datos para un mismo autor y desde un mismo origen de datos, pues forzando a correr nuevamente el proceso inicial se reagruparán los conjuntos de autores y coautores en base a estos nuevos datos.

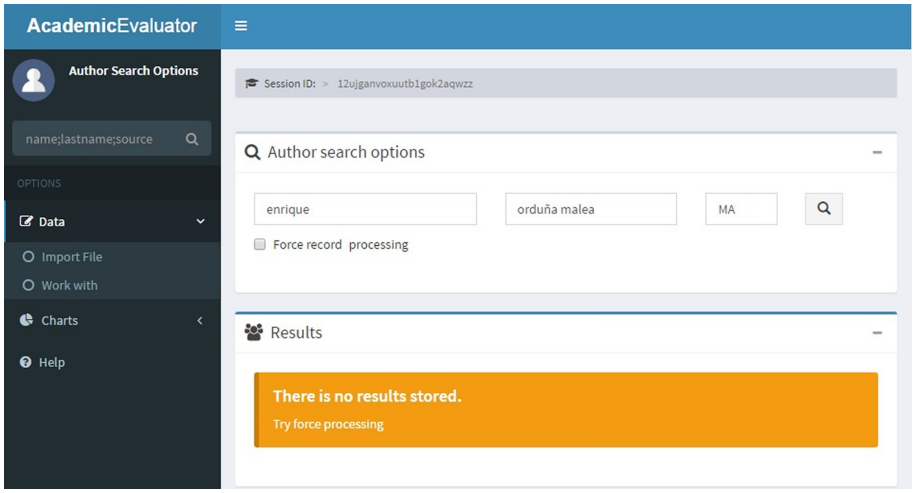


Figura 5.1: Pantalla inicial de búsqueda por autor

Habiendo indicado que se realice el procesamiento, el sistema informa que no existen registros almacenados y ofrece la opción de recuperarlos con la API de *Microsoft Academic* (ver Figura 5.2), al confirmar dicha opción se realiza la recuperación y al finalizar dicho proceso el sistema indica la cantidad de registros recuperados (ver Figura 5.3).

El proceso de recuperación inicia consultando la AK API para extraer los registros buscando por nombre de autor, en este caso, la búsqueda se hace para el autor Enrique Orduña Malea como se indica en Algoritmo 5.1. Para realizar esto se especifica el atributo *AuN* que corresponde al nombre del autor que pertenece a *Author Entity*. Este entidad posee otros atributos que describen al autor, pero en esta recuperación solo se

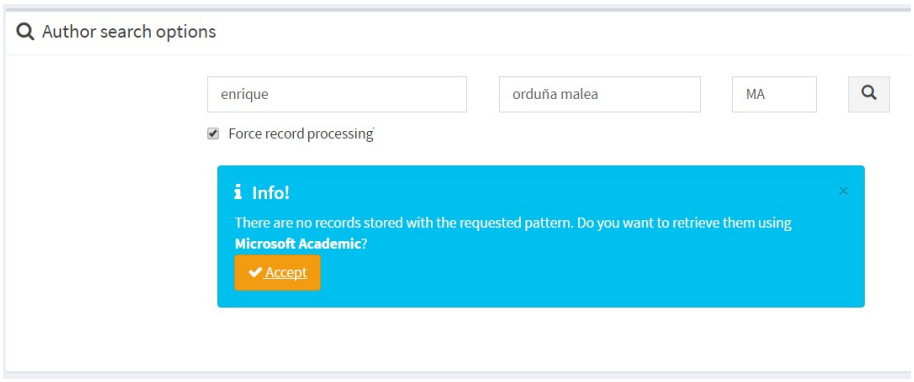


Figura 5.2: Opción de recuperación de registros con *Microsoft Academic*

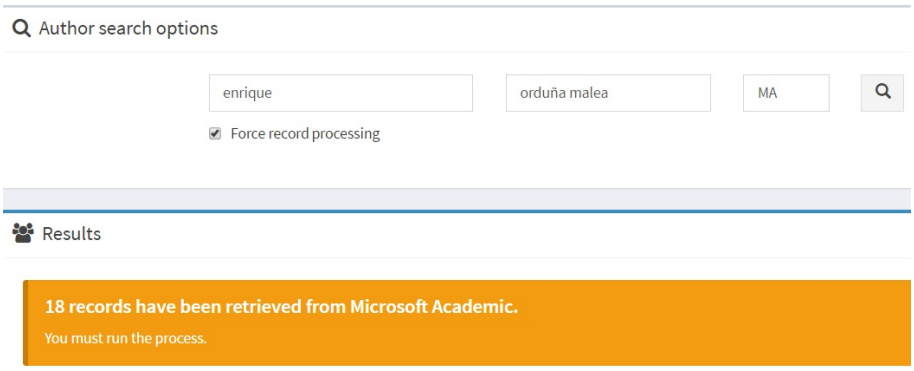


Figura 5.3: Total de registros recuperados con *Microsoft Academic*

utilizará el nombre de autor. La consulta queda expresada en la forma "Composite(AA.AuN==énrique orduna malea)", este valor se asigna al atributo *queryString["expr"]* el cual es el parámetro más importante de la consulta.

El nombre de autor debe estar normalizado, esto es, se intercambian las vocales acentuadas por vocales sin acentuar, se intercambian "ñ" por "n" y se eliminan los signos de puntuación. No es un criterio necesario para realizar las búsquedas por nombre de autor, que esté normalizado el nombre, pero al hacerlo mejoran notablemente los resultados de dicha consulta.

```
WebBrowser.Navigate(URL);
client.DefaultRequestHeaders.Add("Ocp-Apim-Subscription-
Key", APIKeyAcademic);
// Request parameters
queryString["expr"] = "Composite(AA.AuN=='enrique
    orduna malea')";
queryString["model"] = "latest";
queryString["count"] = "1000";
queryString["offset"] = "0";
queryString["orderby"] = "Y:desc";
queryString["attributes"] =
    "Id,Ti,Y,CC,AA.AuN,AA.AuId,AA.AfN,F.FN,J.JN,W,E";
var uri =
    "https://westus.api.cognitive.microsoft.com/academic/
v1.0/evaluate?" + queryString;

var response = await client.GetAsync(uri);
var jsonResponse = await
    response.Content.ReadAsStringAsync();
```

Algoritmos 5.1 Consulta a AK API por nombre de autor

En el experimento que se realizó, es impensable que este autor tan productivo posea solo 18 registros, el caso es que *Microsoft Academic* tiene indexados solo esa cantidad de registros con la forma "enrique orduna malea", pero al variar un poco la forma del nombre, es decir, quitando los espacios en blanco del apellido el resultado es otro (ver Figura 5.4).

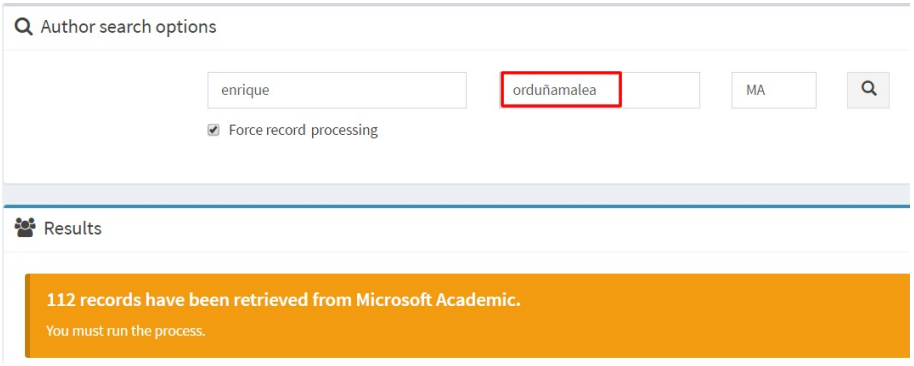


Figura 5.4: Total de registros recuperados con *Microsoft Academic* con una variación del apellido

Esto no es un error de la aplicación, sino un error del motor pues posee a un mismo autor indexado de forma distinta y no realizó la fusión correspondiente o en el mejor de los casos no realizó la desambiguación interna, de manera de mostrar al usuario final un único resultado. Por ello es que internamente, se resolvió este inconveniente, haciendo dos búsquedas, la primera con el nombre de autor normalizado (sin tildes ni acentos, sin "ñ" y sin signos de puntuación) y la segunda con el nombre de autor normalizado y eliminando los espacios en blanco entre los apellidos, ya que se detectó que *Microsoft Academic* los indexa con esas dos variantes, si el autor posee más de un nombre y/o mas de dos apellidos, también se resolvió de recuperar los registros con esas distintas variantes (ver Figura 5.5).

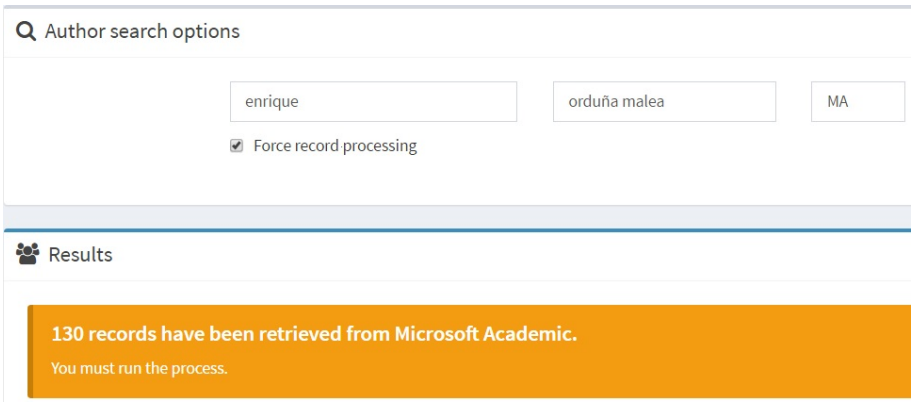


Figura 5.5: Total de registros recuperados con *Microsoft Academic* con ambas variaciones del apellido

5.1.2. Procesamiento inicial

Una vez recolectados los datos, el siguiente paso es correr el proceso como lo indica la alerta entregada por el sistema en la Figura 5.5, este proceso, como se indicó oportunamente, identifica los distintos autores en base a los coautores, es decir, arma las agrupaciones teniendo en cuenta los patrones de colaboración entre los coautores, o dicho de otra manera, los coautores que comparten al menos alguna publicación formarán parte de una misma agrupación. En las Figura 5.6 y en la Tabla 5.1 se exhiben las agrupaciones resultantes.

Figura 5.6: Agrupaciones resultantes Parte 2

Tabla: 5.1: Listado de agrupaciones iniciales

Nombre de autor	**Coautores**
enrique ordunamalea	emilio delgado lopezcozar, alberto martinmartin, juan manuel ayllon, joseantonio ontalbaruiperez, jorge serranocobos, nuria lloret romero, tomas baiget, juan manuel ayllon millan, evaristo jimenezcontreras, rafael ruizperez, javier guallar, daniel torressalinas, alicia selles carot, antonia ferrersapena, fernanda peset, josepmanuel rodriguezgairin, adolfo alonsoarroyo, david azorinricharte, jose antonio ontalba y ruiperez, angeles calduchlosa, annewil harzing, cristina i font, monica antolicalleja, maria vano planells, peter jacso, delgado lopezcozar e, isabel olea, imma subiratscoll, alvaro cabezasclavijo, diego marcos cartagena, daniel roblescrespo, monica caballo, elea gimeneztoledo, jose r perezaguera, silvia redondo, francisco tosete
enrique ordunamalea	mike thelwall, kayvan kousha
enrique ordunamalea	paula torresperez, eva mendezrodriguez
enrique ordunamalea	selenay aytac
enrique ordunamalea	isidro f aguillo, jose luis ortega

enrique orduna malea	margarita cabrera mendez, vicent gimenez chornet
enrique ordunamalea	john j regazzi
enrique ordunamalea	juan antonio pastorsanchez, tomas saorin
enrique ordunamalea	jose miguel carot
enrique ordunamalea	miguel sanchezcervillo
enrique orduna malea	tesina de master
enrique orduna malea	lledo felip vidal

Como se observa, para este autor se encontraron 12 agrupaciones de coautores que al parecer no se conocen. Cada recuadro muestra el nombre del autor y la cantidad de publicaciones a su nombre (por ejemplo: **enrique ordunamalea (104)**), y debajo del nombre el listado de coautores de esa agrupación. La aplicación permite ver las publicaciones de cada autor y coautor, haciendo clic en el número encerrado entre paréntesis (ver Figura 5.7).

List of Publications

Show 10 entries Search:

Title	Authors	Year	CiteNumber	Type
methods for estimating the size of google scholar	enrique ordunamalea,juan manuel ayllon,alberto martinmartin,emilio delgado lopezcozar	2015	23	[PDF]
google scholar metrics evolution an analysis according to languages	enrique ordunamalea,emilio delgado lopezcozar	2014	15	[PDF]
the silent fading of an academic search engine the case of microsoft academic search	enrique ordunamalea,alberto martinmartin,juan manuel ayllon,emilio delgado lopezcozar	2014	13	[PDF]
selective linking from social platforms to university websites a case study of the spanish academic system	enrique ordunamalea,joseantonio ontalbaruiperez	2013	11	[PDF]
presencia y visibilidad web de las universidades publicas espanolas	enrique ordunamalea,jorge serranocobos,jose antonio ontalbaruiperez,nuria lloretromero	2010	10	[PDF]
las universidades publicas espanolas en google scholar presencia y evolucion de su publicacion academica web	enrique ordunamalea,jorge serranocobos,nuria lloretromero	2009	8	[PDF]
journal scholar una alternativa internacional gratuita y de libre acceso para medir el impacto de las revistas de arte humanidades y ciencias sociales	emilio delgado lopezcozar,enrique ordunamalea,diego marcos cartagena,evaristo jimenezcontreras,rafael ruizperez	2012	8	[PDF]

Figura 5.7: Listado de publicaciones de un autor

Estas agrupaciones se formaron solamente revisando los datos de los autores de cada publicación, para comprobar si el nombre del autor se encuentra entre estos nombres, sino es así el registro se marca para un reproceso o segunda comprobación. Por el contrario si el autor si se encuentra entre los autores que firman la publicación, se revisa en la base

de datos de la aplicación, si no es un autor previamente almacenado (se consultan las tablas *Authors* y *VariantsAuthors*) y además se revisa si alguno de los coautores del registro almacenado no se encuentra almacenado en la base de datos (se consultan las tablas *CoAuthors* y *VariantsCoAuthors*). Si no se encuentra ningún coautor se crea una nueva agrupación, cada agrupación tiene al nombre del autor o su variante a la cabeza, y dependiendo de este el listado de coautores. Si una persona conoce al autor buscado, podrá observar que algunas agrupaciones están demás, es decir, pues el autor de ambas agrupaciones es la misma persona, pero este hecho la aplicación lo desconoce pues se limita a los datos con que cuenta, en este caso los nombres de autores de las publicaciones. En los procesos posteriores se intentará hallar alguna relación entre estas agrupaciones de manera de reducir la cantidad de estas.

5.1.3. Reproceso

El reproceso es un mecanismo que permite identificar aquellos registros que no pudieron procesarse en la etapa inicial, de existir, el sistema ofrecerá un link para observar cuáles son. Estos registros son aquellos que no pudieron encajar en ningún grupo de coautores (pues los coautores de estos registros no son los ya identificados) y además el nombre del autor o alguna de sus variantes no se encontró en el listado de autores de la publicación. Para el caso que se está tratando, no existen registros que requieran un reproceso puesto que el nombre del autor buscado sí se encontró en los 130 registros recuperados, aunque no siempre sucede de este modo. Esta opción es especialmente útil para los registros de *Google Scholar* los cuales poseen para diversas formas para el mismo nombre de autor.

Microsoft Academic posee los nombres de autor mucho mas normalizados que *Google Scholar*, es más, por cada publicación almacenada en este motor, existe un listado de autores y no una simple área destinada para ello como lo ofrece *Google Scholar*. Por esta razón, al recuperar los registros de *Microsoft Academic*, el nombre de autor o alguna de sus variantes sí está presente en el listado de autores de la publicación, es por esto que no existen registros a reprocesar para este experimento.

5.1.4. Análisis de transitividad

Una vez finalizado el "reproceso", inicia un proceso que lleva por nombre "análisis de transitividad" puesto que estudia la existencia de alguna relación entre dos grupos de autores. La idea es reagrupar los grupos de autores formados anteriormente, infiriendo alguna relación no explicita que no se puede hallar con los datos recolectados. Entiéndase por relación a: la existencia de al menos una publicación no indexada por el motor académico y que relaciona al menos a dos coautores de dos grupos distintos, o valiéndose de la popularidad de los resultados ofrecidos por el motor de búsqueda de propósito general *Bing*, revisar si del top 10 de resultados existe alguna relación entre los coautores más productivos de ambos grupos o entre el coautor más productivo de un grupo y cualquier coautor del otro grupo. Esta relación que se intenta hallar no siempre viene en forma de una publicación científica, por ello la aclaración del uso del motor *Bing*, debido a que las respuestas entregadas por este motor son las páginas web indexadas que respetan los criterios de búsqueda solicitados. Vale también aclarar que no se utiliza el motor de búsqueda *Google* (que es más potente y posee una mayor cobertura) a causa de que este motor posee los mismos bloqueos impuestos a *Google Scholar* al momento de realizar búsquedas automáticas.

El proceso en cuestión se inicia ordenando todos los grupos de autores encontrados (cada grupo posee un conjunto de coautores) de forma descendente por cantidad de publicaciones, esto se realiza consultando la tabla *Authors*, extrayendo el identificador y nombre de cada autor (ver Figura 5.8a). Con este listado se arman las combinaciones de autores de dos en dos que serán evaluadas, cada combinación representa un registro dentro de la tabla *Combinations*, esta tabla almacena el número de recorrido (campo *Round*), identificador del autor de ambos grupos (campos *AuthorId1* y *AuthorId2*) y el estado de la comprobación (campo *State*). La Figura 5.8b ofrece el esquema de combinaciones para la primer ronda o ronda inicial.

El procedimiento continúa tomando de cada combinación los identificadores de autores, para obtener el listado de coautores de cada uno de los dos grupos ($Group_1$ y $Group_2$). Ambos grupos son ordenados de forma descendente por la cantidad de publicaciones de los coautores, luego se toma el coautor más productivo del $Group_1$ para buscar si está relacionado con al menos alguno de los dos coautores más productivos del

AuthorId	AuthorName
112	enrique ordunamalea
118	enrique ordunamalea
117	enrique ordunamalea
120	enrique ordunamalea
126	enrique orduna malea
121	enrique ordunamalea
122	enrique ordunamalea
123	enrique ordunamalea
125	enrique ordunamalea
113	enrique orduna malea
115	enrique ordunamalea
116	enrique orduna malea

(a) Listado de autores

Round	AuthorId1	AuthorId2	Estate
1	112	118	
1	117	120	
1	126	121	
1	122	123	
1	125	113	
1	115	116	

(b) Listado de combinaciones

Figura 5.8: Armado de combinaciones a partir del listado autores

$Group_2$. La Figura 5.9 expone lo mencionado.

Grupo 1	
Autor	#Publicaciones
enrique ordunamalea	104
Coautor	#Publicaciones
emilio delgado lopezcozar	52
alberto martinmartin	30
juan manuel ayllon	16
joseantonio ontalbaruiperez	13
jorge serranocobos	9

Grupo 2	
Autor	#Publicaciones
enrique ordunamalea	4
Coautor	#Publicaciones
isidro f aguillo	4
jose luis ortega	2

Figura 5.9: Evaluación de dos grupos de autores

Si se logra hallar que ambos grupos están relacionados, estos se fusionan, el $Group_2$ desaparece y tanto los coautores como las publicaciones de cada uno se actualizan para indicar la nueva referencia hacia el $Group_1$, en este sentido, el $Group_1$ ahora es un grupo más grande pues posee los elementos de ambos grupos. Como la combinación evaluada fue satisfactoria, la misma es actualizada al estado "OK" ($State = OK$), luego esta combinación fusionada en un mismo grupo será candidata para la segunda vuelta de evaluaciones con las restantes combinaciones satisfactorias. Si la combinación evaluada no fue satisfactoria, el estado de la misma se actualiza a "NO" ($State = NO$). En las rondas o vueltas suce-

sivas se toman solamente las combinaciones marcadas como "OK" para generar un nuevo registro de combinación a evaluar, el procedimiento de evaluación es idéntico al ejecutado en la ronda inicial. La Figura 5.10 muestra el estado de las combinaciones luego de finalizada la ronda final, para este ejemplo solo hubo dos rondas, pues en la primera solo una combinación fue marcada como "OK", y al no poseer otra del mismo estado, las rondas de evaluación finalizan.

Round	AuthorId1	AuthorId2	Estate
1	112	118	NO
1	117	120	OK
1	126	121	NO
1	122	123	NO
1	125	113	NO
1	115	116	NO
2	117		OK

Figura 5.10: Listado de combinaciones luego de la ronda final

Luego de haber evaluado todas las combinaciones y no habiendo más rondas posibles por generar, se toman todas las combinaciones de todas las rondas ejecutadas marcadas como "NO", de estas se recuperan los identificadores de autores y se arma un listado de autores, por cada elemento de este listado se obtiene el conjunto de coautores ordenados de forma descendente por cantidad de publicaciones, se toma al coautor más productivo del $Group_n$ y a los 15 coautores más productivos de todas las demás agrupaciones (no se incluyen los coautores de $Group_n$), estos 15 coautores se ordenan de forma descendente por cantidad de publicaciones (ver Figura 5.11). Luego se realiza un consulta a la *Bing Web Search API* con el nombre de autor analizado y el nombre de autor más productivo del $Group_n$, por cada resultado devuelto se toma la URL y se recupera el recurso apuntado, se extrae el texto y se busca en los primeros 10 resultados la existencia de al menos uno de los 15 coautores seleccionados, con la primer coincidencia se toma el grupo al que pertenece el coautor encontrado y se fusiona con el $Group_n$.

Luego de finalizado este proceso, el sistema mostrará las agrupaciones resultantes (ver Figura 5.12). Dicho resultado depende en gran medida de las combinaciones de grupos de coautores tomadas para realizar las comprobaciones, teniendo en cuenta que realizar comprobaciones "todos

Grupo 1	
Autor	#Publicaciones
enrique ordunamalea	104
Coautor	#Publicaciones
emilio delgado lopezcozar	52
alberto martinmartin	30
juan manuel ayllon	16
joseantonio ontalbaruiperez	13
jorge serranocobos	9

El resto de los grupos		
Coautor	#Publicaciones	Grupo
isidro f aguillo	4	2
john j regazzi	3	4
jose luis ortega	2	2
lledo felip vidal	2	5
margarita cabrera mend	1	3
vicent gimenez chornet	1	3
juan antonio pastorsanc	1	6
tomas saorin	1	6
jose miguel carot	1	7
miguel sanchezcervillo	1	8
tesina de master	1	9
mike thelwall	1	10
kayvan kousha	1	10
paula torresperez	1	11
eva mendezrodriguez	1	11

Figura 5.11: Evaluación de un grupo contra los coautores del resto de los grupos

contra todos" es impracticable, por ello se seleccionan los dos coautores más productivos de cada grupo ponderando a estos en contraposición a los coautores menos productivos, bajo el supuesto que es más probable que estos coautores se conozcan o posean intereses en común.

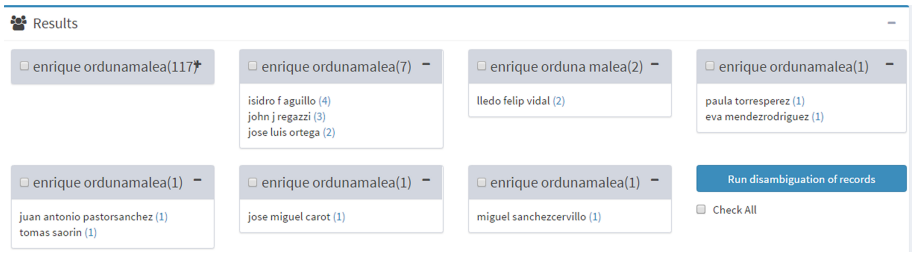


Figura 5.12: Agrupaciones resultantes luego del análisis de transitividad

Se realizó la división de todos estos procesos para que el lector comprenda el ámbito de aplicación de cada uno de ellos, pero en realidad el sistema ejecuta los tres procesos, el proceso inicial, el reproceso y análisis de transitividad en un solo conjunto de modo transparente, ofreciendo los resultados automáticamente sin que el usuario observe la transición entre uno y otro.

5.1.5. Detección de duplicados, desambiguación de registros

Luego de obtener las agrupaciones resultantes, el sistema ofrece la opción de lanzar el proceso de desambiguación de registros (*Run disambiguation of records*). Dicho proceso requiere la selección de al menos una agrupación de coautores para proceder con dicho análisis, la opción "Check All" selecciona todos los grupos mostrados. En este ejemplo se eligieron las dos agrupaciones con mayor número de registros (117 y 7 registros).

La detección de duplicados parte de un conjunto de publicaciones pertenecientes a uno o más autores elegidos por el usuario, esto es así pues es el usuario de la aplicación quien decide reagrupar los autores según lo considere, por ejemplo, si el usuario conoce al autor analizado y existen agrupaciones que el sistema no pudo reagrupar, el puede decidir elegirlas para el análisis, de igual modo, si el usuario no conoce al autor puede elegir distintas combinaciones de agrupaciones o todas, y ver el comportamiento de los resultados. Esta flexibilidad puede ser bastante útil, ya que llegados a este punto, la demora de los procesos para la desambiguación de publicaciones y cálculo de resultados, es de solo algunos segundos.

El gran costo que demanda hacer comparaciones de todos contra todos obliga a escoger algún esquema de agrupamiento para aplicar alguna técnica de detección de duplicados. En este trabajo se eligió crear conjuntos de publicaciones por año de publicación, y sobre los elementos de estos conjuntos evaluar las publicaciones tomadas de dos en dos en un esquema de todos contra todos como el que se presenta en Algoritmo 5.2.

```
var lstYears = listR.Select(o => o.Year).Distinct();
    //del listado de registros, se obtienen los años
    únicos
foreach (int y in lstYears) //se recorre el listado de
    años de publicación
{
  var lstRecords = GetRecordByYear(y).ToList();
  for (int i = 0; i < lstRecords.Count - 1; i++)
  {
    for (int j = i + 1; j < lstRecords.Count; j++)
    {
```

```
      var dAux =
          JaroWinklerTFIDF.score(lstRecords[i].Title,
          lstRecords[j].Title);
      if(dAux >= TitleThreshold)
      {
        InsertDuplicate(lstRecords[i].RecordId,
            lstRecords[j].RecordId);
      }
    }
  }
}
```

Algoritmos 5.2 Detección de duplicados

La variable *TitleThreshold* es el umbral definido para comparar dos títulos de publicaciones, el valor de la misma se fijó en 0.92. Si el resultado de aplicar la función de distancia de edición *JaroWinklerTFIDF* retorna un valor mayor o igual a 0.92, se concluye que ambas publicaciones son posibles duplicados. Como se vio en el **Capítulo 3**, se evaluaron otras funciones de distancia de edición, pero se optó por usar esta función por la adaptación al comparar títulos de publicaciones.

Se estudió utilizar otros esquemas o técnicas, por ejemplo la utilización de la función *Soundex*[4] incorporada en el motor de bases de datos utilizado en el proyecto. Pero esta función es impracticable para un conjunto de términos como el título de una publicación, ya que el uso normal es para términos individuales, por otro lado esta función es muy sensible a errores ortotipográficos.

Los términos desempeñan diferentes funciones en los textos. En términos generales, los sustantivos, los verbos y los adjetivos son más discriminatorios que los adverbios, los conectivos, los pronombres y los números. Es incorrecto asignar un mismo peso a todos los términos. Dado que pocos términos se producirán más de una vez en un solo texto corto, el esquema TF-IDF tradicional es inapropiado para textos cortos (Washio et al., 2008).

El resultado de la detección de registros duplicados se observa en la Tabla 5.2.

[4] https://docs.microsoft.com/en-us/sql/t-sql/functions/soundex-transact-sql

Tabla: 5.2: Listado de registros duplicados resultante del análisis

#	**Título**	**Año**	**Citas**
1	la bibliometria que viene almetrics author level metrics y las multiples caras del impacto de un autor	2016	0
2	la bibliometria que viene almetrics author level metrics y las multiples caras del impacto de un autor	2016	0
3	redes de conectividad entre empresas tecnologicas a traves de un analisis metrico longitudinal de menciones de usuario en twitter	2016	0
4	redes de conectividad entre empresas tecnologicas a traves de un analisis metrico longitudinal de menciones de usuario en twitter	2016	0
5	h index scholar el indice h de los profesores de las universidades publicas espanolas en humanidades y ciencias sociales	2014	4
6	h index scholar el indice h de los profesores de las universidades publicas espanolas en humanidades y ciencias sociales	2014	0
7	aggregation of the web performance of internal university units as a method of quantitative analysis of a university system the case of spain	2013	3
8	aggregation of the web performance of internal university units as a method of quantitative analysis of a university system the case of spain	2013	1
9	aggregation of the web performance of internal university units as a method of quantitative analysis of a university system the case of spain aggregation of the web performance of internal university units as a method of quantitative analysis of a university system the case of spain	2013	0
10	impacto de los repositorios a traves de tecnicas cibermetricas el caso general de latinoamerica y especial de costa rica	2013	0
11	impacto de los repositorios a traves de tecnicas cibermetricas el caso general de latinoamerica y especial de costa rica material complementario	2013	0
12	fuentes de enlaces web para analisis cibermetricos 2012	2012	1
13	fuentes de enlaces web para analisis cibermetricos	2012	0
14	graphic multimedia and blog content presence in the spanish academic web space	2012	3
15	graphic multimedia and blog content presence in the spanish academic web space	2012	1
16	personalizacion e interactividad en los rankings de universidades publicados en la web	2011	4
17	personalizacion e interactividad en los rankings de universiades publicados en la web	2011	0

18	analisis de la correlacion entre la audiencia web de los medios digitales de prensa espanola y su visibilidad en gestores sociales en noticias	2010	0
19	analisis de la correlacion entre la audiencia web de los medios digitales de prensa espanola y su visibilidad en gestores sociales en noticias	2010	0
20	proposal of a goal oriented shared catalog model	2010	1
21	proposal of a goal oriented shared catalog model	2010	0
22	ranking de universidades en la union europea aproximacion multidimensional a una realidad compleja	2010	0
23	ranking de universidades en la union europea aproximacion multidimensional a una realidad compleja	2010	0
24	attention profile informacion a la que prestamos atencion	2009	0
25	attention profile informacion a la que prestamos atencion	2009	0
26	el proyecto visado arquitectnico descripcin caracterizacin y normalizacin documental	2009	0
27	el proyecto visado arquitectonico descripcion caracterizacion y normalizacion documental	2009	0
28	propuesta de indicadores metricos para gestores sociales de noticias analisis de la prensa digital espanola en meneame	2009	1
29	propuesta de indicadores metricos para gestores sociales de noticias analisis de la prensa digital espanola en meneame	2009	0
30	reutilizacion e intercambio de objetos digitales compuestos en la web el proyecto oai ore	2009	0
31	reutilizacion e intercambio de objetos digitales compuestos en la web el proyecto oai ore	2009	0
32	analisis de los frbr en la ejecucion de tareas genericas en catalogos compartidos	2008	0
33	analisis de los frbr en la ejecucion de tareas genericas en catalogos compartidos	2008	0

Las filas de color azul indican al registro que se seleccionó como principal, mientras que el o los registros debajo de color blanco son los duplicados. En la fila 7 se observa un registro con 2 copias adicionales (filas 8 y 9), este ejemplo es interesante porque el duplicado (fila 8) posee una cita, como el origen de datos no entrega los registros de citas, no es posible realizar la comprobación y fusión de las mismas, por esta razón se toma el registro con mayor número de citas y el resto como duplicados.

Una vez finalizado el proceso de desambiguación de registros, la aplicación ofrece un panel de resultados como el mostrado en la Figura 5.13.

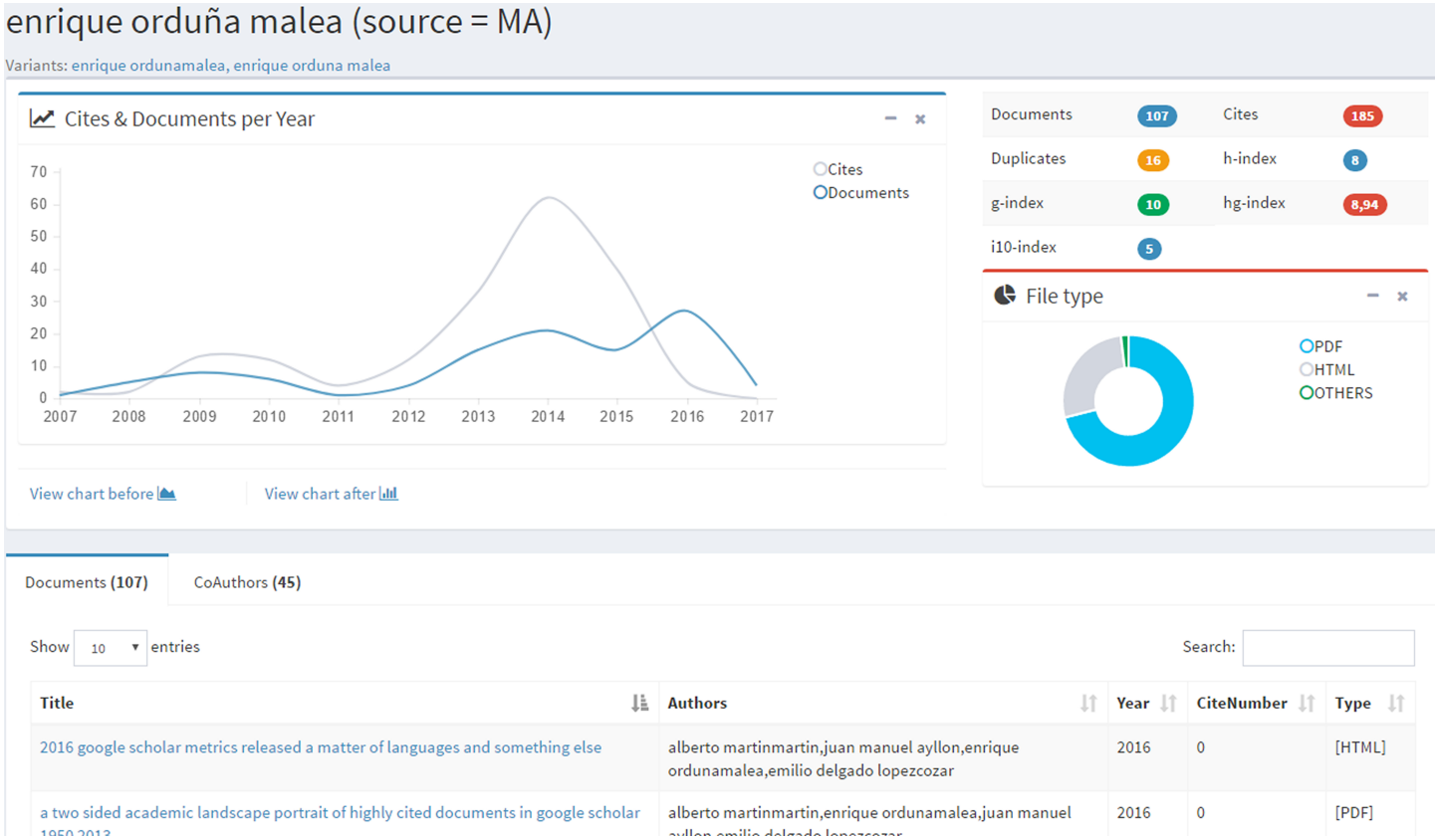


Figura 5.13: Panel de resultados para el autor Enrique Orduña Malea, *source*=MA

Este panel posee varias secciones de resultados. En la parte superior indica el nombre del autor analizado, el origen de datos seleccionado y las variantes del nombre encontradas. Una de las secciones de este panel muestra un gráfico de lineas con dos curvas, la cantidad de citas recibidas (curva de color gris) y la cantidad de documentos publicados (curva de color azul) por año de publicación (ver Figura 5.14). Este gráfico es sencillo y presenta como fluctúan estas dos variables. Abajo de este gráfico se observan dos opciones, *View chart before* y *View chart after*, ambas serán tratadas en el apartado siguiente.

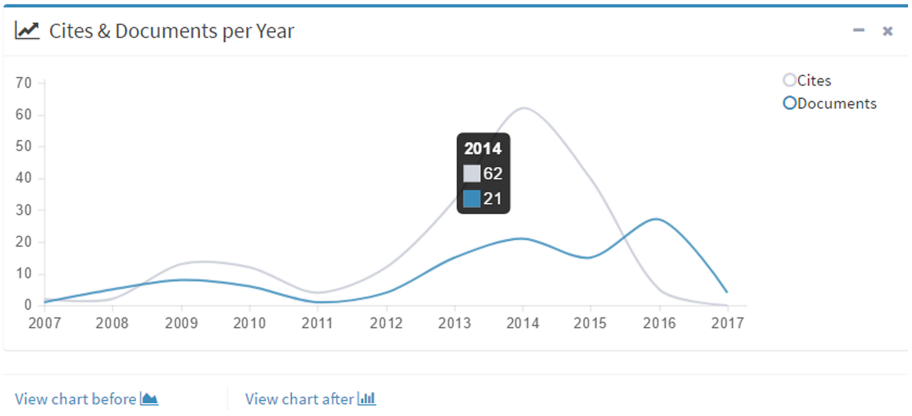


Figura 5.14: Panel de *Cites & Documents per Year*

Dentro de este panel existe un cuadro resumen con los indicadores propuestos oportunamente (ver Figura 5.15).

Documents	107	Cites	185
Duplicates	16	h-index	8
g-index	10	hg-index	8,94
i10-index	5		

Figura 5.15: Recuadro con resumen de indicadores

Los indicadores presentados son:

- *Documents*: cantidad de documentos publicados sin tener en cuenta los registros duplicados.
- *Cites*: cantidad de citas recibidas por estos documentos, para el caso de los duplicados, lo que se hizo fue fusionar las citas recibidas en caso de haber recolectado las mismas, caso contrario se toma el registro con mayor número de citas.
- *Duplicates*: cantidad de documentos duplicados. Siguiendo el ejemplo, los documentos seleccionados sumaban 124 puesto que se habían elegido las agrupaciones que poseían 117 y 7 documentos. Al sumar la cantidad de *Documents (107)* y *Duplicates (16)* el total es 123. Esto no es un error y se debe a que un registro se encuentra repetido 3 veces, como solo se contabiliza un documento original y un documento duplicado, esto explica la diferencia en los valores indicados. El sistema ofrece la opción de ver los documentos duplicados haciendo clic en el valor indicado en *Duplicates*, este es un enlace que abre una ventana con los elementos identificados como posibles duplicados, la Figura 5.16 exhibe estos elementos, como se observa, en el recuadro superior (en color rojo) se encuentra el registro que posee 3 copias y en el recuadro de abajo (en color azul) un registro con dos copias.
- h-index: el valor del índice h para el autor analizado.
- g-index: el valor del índice g para el autor analizado.
- hg-index: el valor del índice hg para el autor analizado.
- $i10$-index: el valor del índice $i10$ para el autor analizado.

Continuando con el panel inicial de resultados, debajo del recuadro de indicadores existe un gráfico de dona para los formatos de las publicaciones. Cada tipo de publicación posee un color y un área dentro de la gráfica, al pasar el mouse por cualquier de estas se observa la cantidad de dicho elemento (ver Figura 5.17).

Por último, en este panel se presentan dos solapas (Figura 5.18), una conteniendo el listado de publicaciones resultantes (ver Figura 5.18a) y sobre las que se basan los resultados indicados previamente, y otra solapa

List of Publications

Show 10 entries Search:

Title	Authors	Year	CiteNumber	Type
aggregation of the web performance of internal university units as a method of quantitative analysis of a university system the case of spain	enrique ordunamalea	2013	3	[HTML]
aggregation of the web performance of internal university units as a method of quantitative analysis of a university system the case of spain	enrique orduna malea	2013	1	[HTML]
aggregation of the web performance of internal university units as a method of quantitative analysis of a university system the case of spain aggregation of the web performance of internal university units as a method of quantitative analysis of a university system the case of spain	enrique ordunamalea	2013	0	[OTHERS]
analisis de la correlacion entre la audiencia web de los medios digitales de prensa espanola y su visibilidad en gestores sociales en noticias	enrique orduna malea	2010	0	[HTML]
analisis de la correlacion entre la audiencia web de los medios digitales de prensa espanola y su visibilidad en gestores sociales en noticias	enrique ordunamalea	2010	0	[PDF]

Figura 5.16: Listado de publicaciones duplicadas

con el listando completo de coautores de las agrupaciones de autores seleccionada inicialmente (ver Figura 5.18b). La aplicación permite abrir el listado de publicaciones en coautoría para cada coautor haciendo clic en el nombre de cada persona.

5.1.6. Visualización de resultados

El panel de resultados ya mencionado, posee dos opciones, *View chart before* y *View chart after*, estas se describen en este apartado dedicado especialmente a ellas. Se hace esta división dado que son dos visualizaciones y no simples gráficos como los mostrados previamente. La visualización aquí presentada se desarrolló exclusivamente para este trabajo en un intento de representar el mayor número de variables y dimensiones del problema en cuestión.

En los apartados correspondientes se mencionó la necesidad de contar con una visualización como ésta y las carencias que poseen las herramientas actuales de evaluación científico-académica en este ámbito, en respuesta a ello, la visualización aquí desarrollada viene a proponer

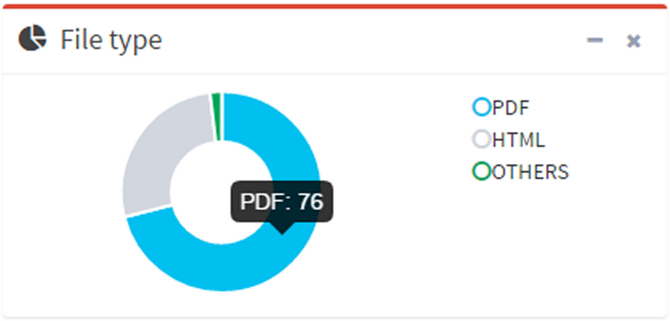
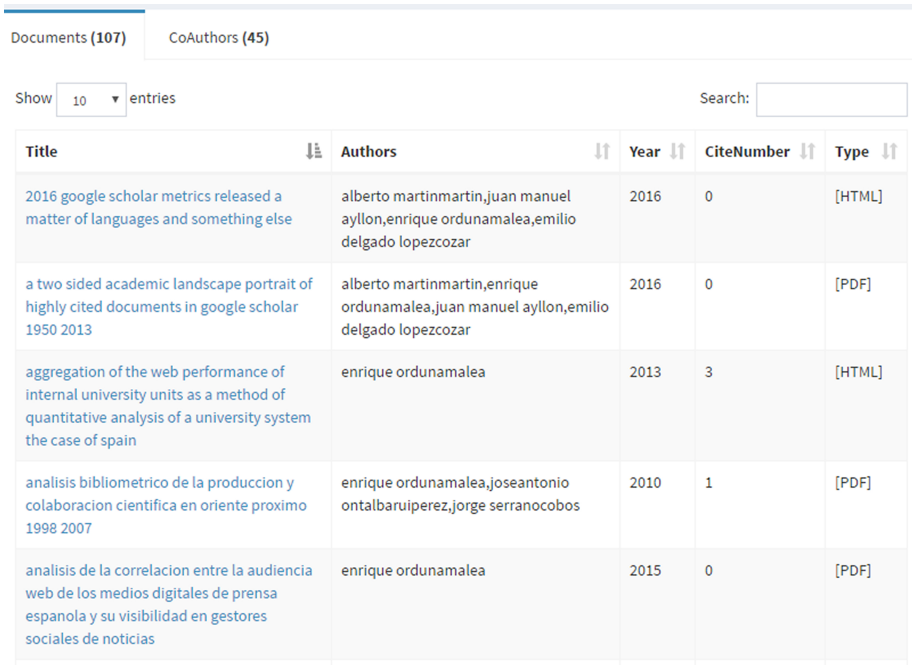


Figura 5.17: Gráfico para los tipos de archivos de las publicaciones

una alternativa que engloba la mayor parte, o al menos las más importantes, de las variables involucradas en un análisis bibliométrico. Esta herramienta, si bien sencilla en su representación, ofrece cierto grado de interacción aprovechable y capaz de otorgar una visión distinta que permita obtener cierto conocimiento a través de un único gráfico.

La opción *View chart before* ofrece la visualización de todas las publicaciones recuperadas para el autor y origen de datos a analizar, es decir, es el conjunto de datos tal cual fue importado o recolectado, sin ningún tipo de control ni filtro (ver Figura 5.19), por su parte la opción *View chart after* utiliza esta misma representación pero con los datos obtenidos luego de todo el proceso de control, revisión y filtrado implementados (ver Figura 5.20).

Comparando ambas figuras se pueden ver algunos cambios entre una y otra, por ejemplo antes del análisis existían dos elementos *Publication source* que no aparecen en la gráfica una vez finalizado el proceso, estos elementos son: *j acad libr* e *internet*. Además entre la visualización inicial y la final existen 23 publicaciones de diferencia. Si bien en las Figuras 5.21 y 5.22 no se aprecian a simple vista estas 23 publicaciones de diferencia, esto responde a la superposición de los elementos, pues 18 poseen 0 citas y varias se encuentran en el mismo año de publicación y con las misma cantidad de coautores. Sin embargo al utilizar los demás filtros de la aplicación, indicando por ejemplo que se muestren solo las publicaciones en formato HTML, el resultado es como se ve en la Figura 5.23.



Documents (107) CoAuthors (45)

Show 10 entries Search:

Title	Authors	Year	CiteNumber	Type
2016 google scholar metrics released a matter of languages and something else	alberto martinmartin,juan manuel ayllon,enrique ordunamalea,emilio delgado lopezcozar	2016	0	[HTML]
a two sided academic landscape portrait of highly cited documents in google scholar 1950 2013	alberto martinmartin,enrique ordunamalea,juan manuel ayllon,emilio delgado lopezcozar	2016	0	[PDF]
aggregation of the web performance of internal university units as a method of quantitative analysis of a university system the case of spain	enrique ordunamalea	2013	3	[HTML]
analisis bibliometrico de la produccion y colaboracion cientifica en oriente proximo 1998 2007	enrique ordunamalea,joseantonio ontalbaruiperez,jorge serranocobos	2010	1	[PDF]
analisis de la correlacion entre la audiencia web de los medios digitales de prensa espanola y su visibilidad en gestores sociales de noticias	enrique ordunamalea	2015	0	[PDF]

(a) Listado de publicaciones resultantes del análisis

(b) Listado de coautores de las publicaciones resultantes

Figura 5.18: Solapas contenedoras con los documentos y coautores de las publicaciones resultantes

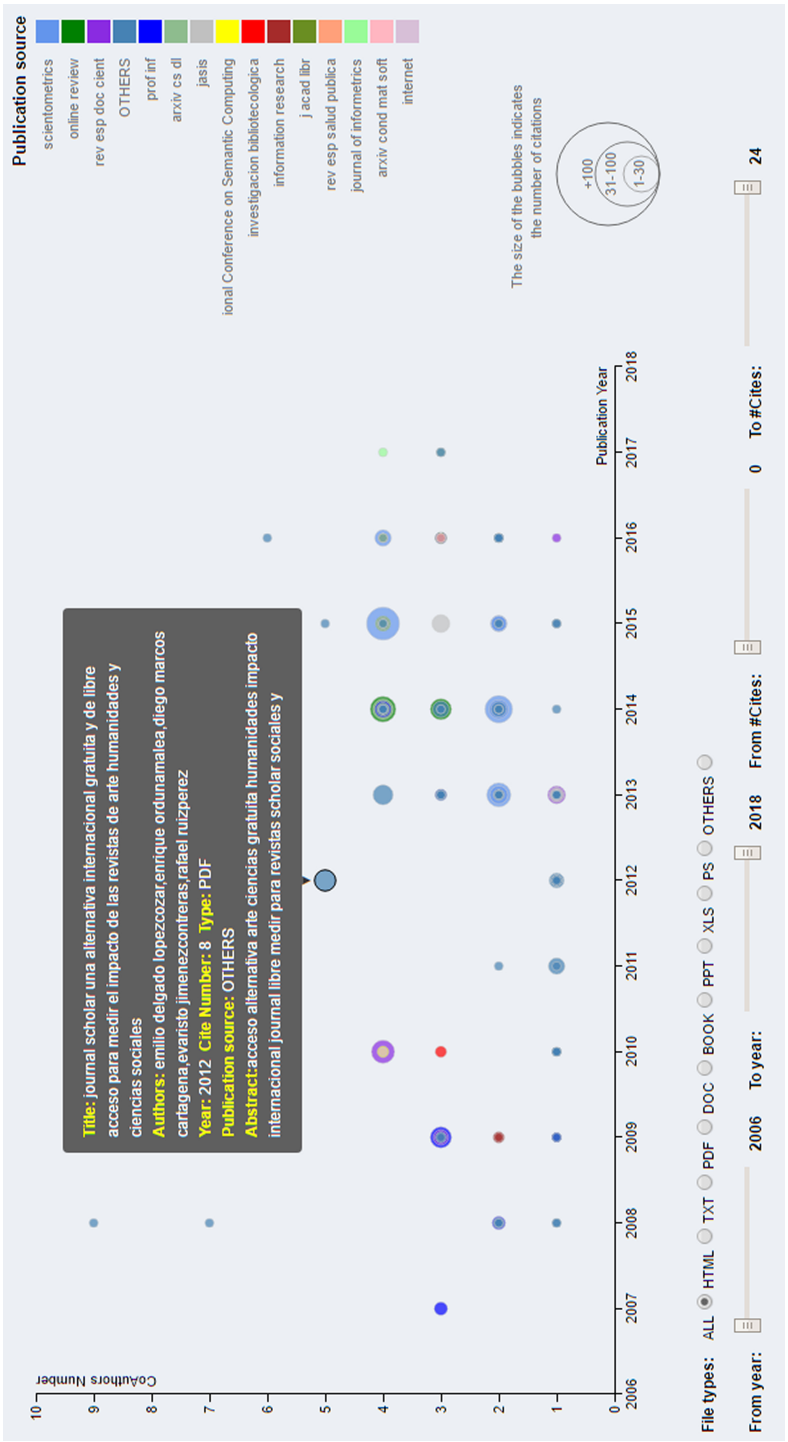


Figura 5.19: Visualización de *scatterplot* antes del análisis (MA)

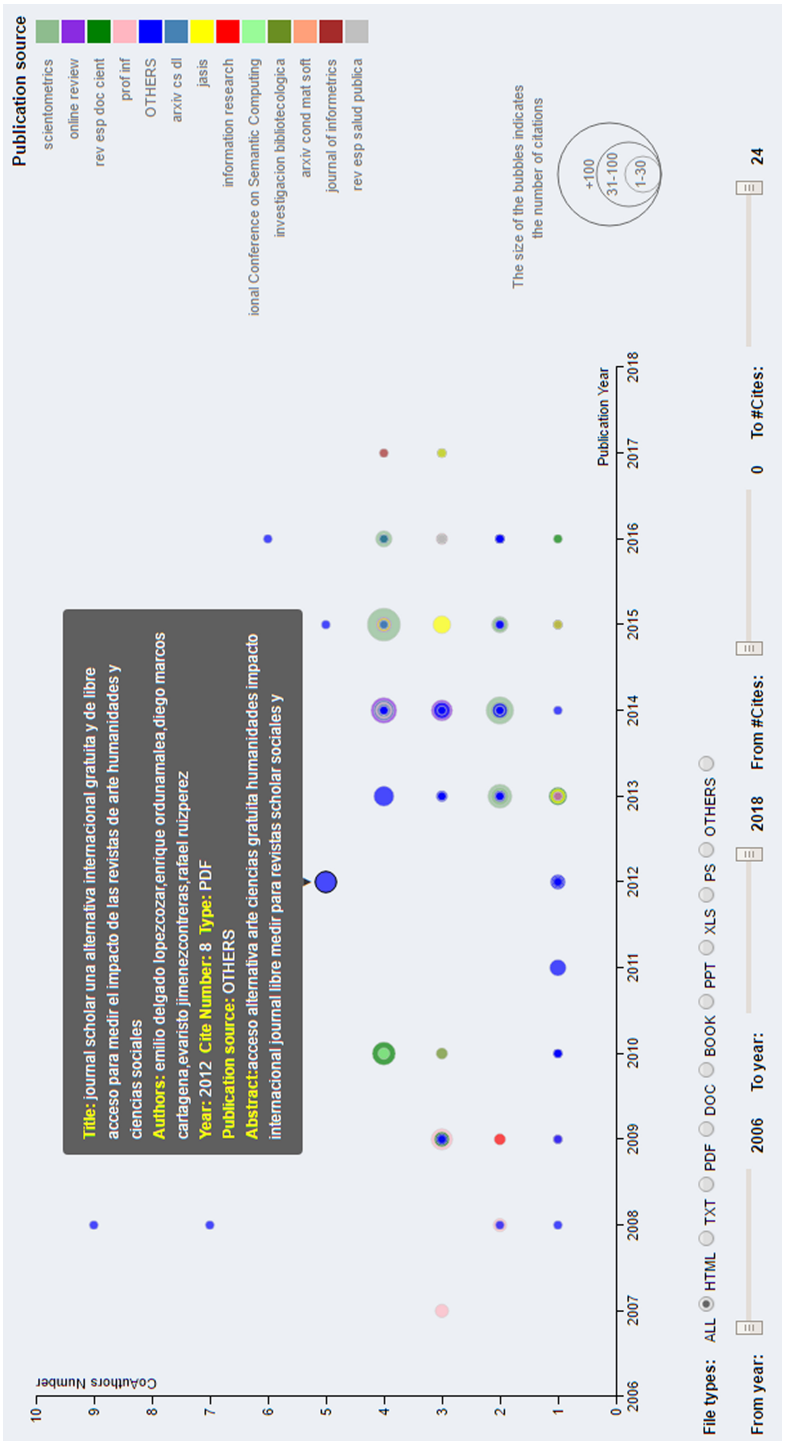


Figura 5.20: Visualización de *scatterplot* luego del análisis (MA)

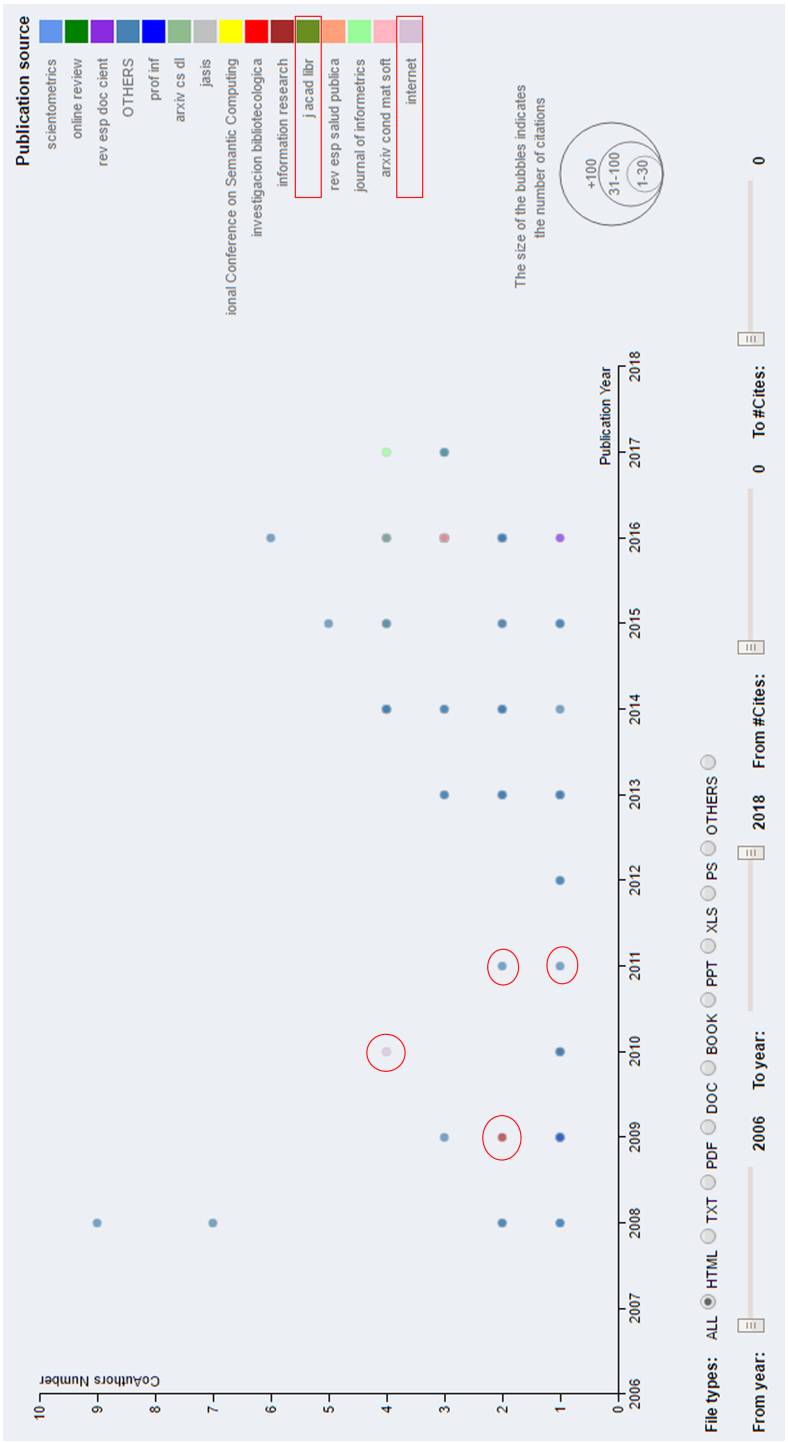


Figura 5.21: Visualización antes del análisis con filtro de 0 número de citas

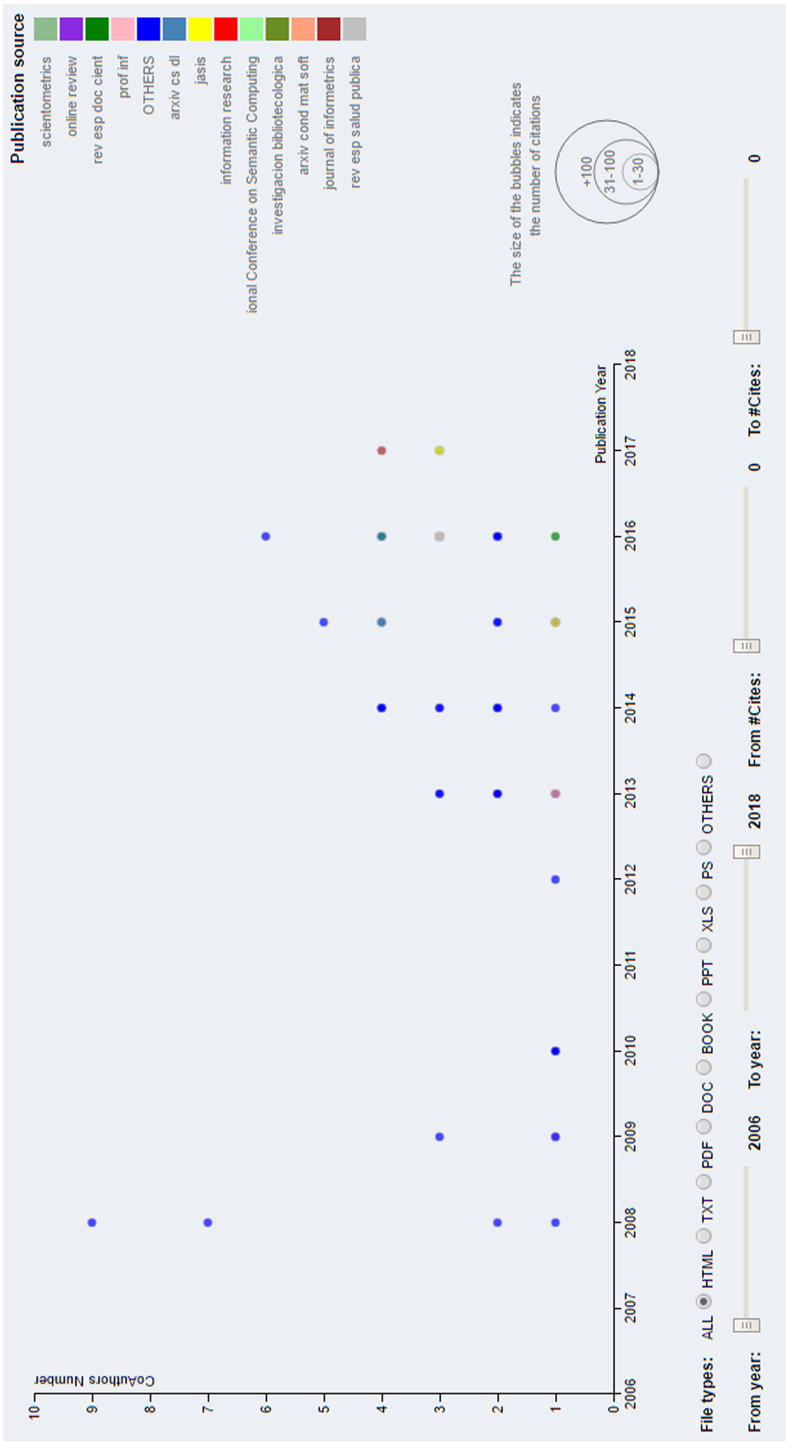


Figura 5.22: Visualización luego del análisis con filtro de 0 número de citas

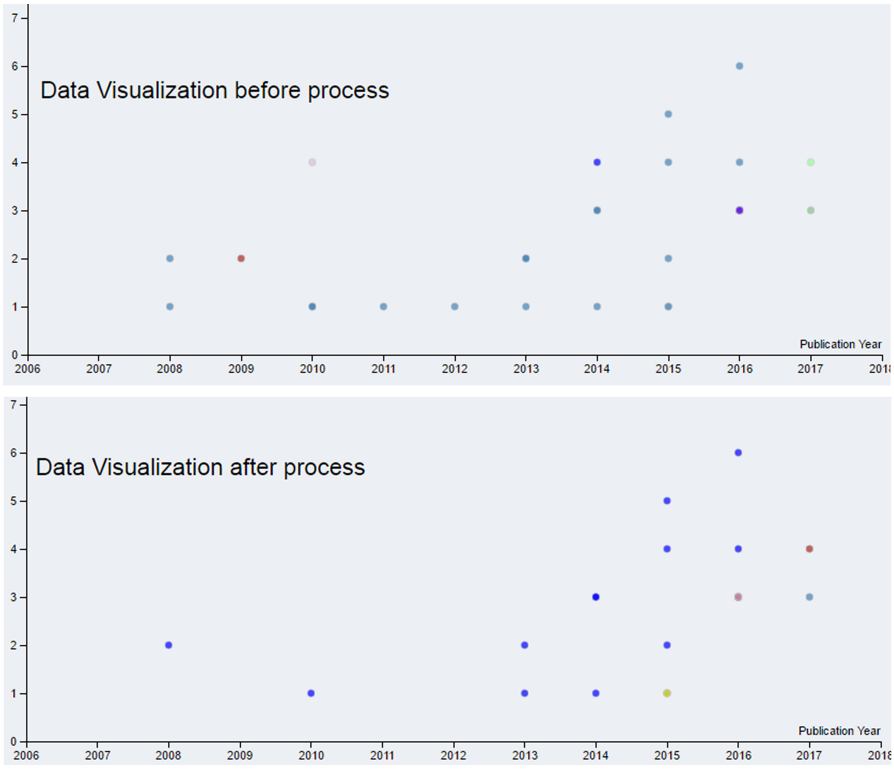


Figura 5.23: Diferencias entre las visualización antes y después del análisis

5.2. Experimentos con *Google Scholar*

Se describe a continuación la serie de pasos para recuperar, procesar y visualizar los datos desde el conjunto de publicaciones de un autor tomando como origen de datos a *Google Scholar*.

5.2.1. Recuperación de información

En el **Capítulo 2** se habló del proceso de recuperación de *Google Scholar* y de los múltiples bloqueos impuestos a las tareas de recolección automática, por ello se decidió no incluir esta recuperación en la herra-

mienta final. Sin embargo la recuperación si se pudo realizar de forma semi-automática.

Como se mencionara previamente, consta de dos pasos, por un lado se realiza la recuperación de los registros de publicaciones y luego de las citas de cada uno de los registros. Como *Google Scholar* no posee una API que permita la recolección automática de registros, las tareas de recolección se basan en recuperar el cuerpo HTML de las respuestas a la URL del navegador, indicando previamente los parámetros de la búsqueda como ser: búsqueda por autor, no incluir citas y no incluir patentes. El proceso de recolección automática fue bloqueado debido a la gran cantidad de datos recolectados, tantos los registros de publicaciones de autores como las citas recibidas por estos. Para poder solucionar este inconveniente, se desarrolló una aplicación de escritorio que implementa los mismos procedimientos antes diseñados para recolectar los datos de esta fuente.

La recolección se realiza cargando la URL de la búsqueda, la misma contiene el nombre del autor a recuperar:

```
https://scholar.google.com.ar/scholar?as_vis=1&q=autor:daniel+
autor:torres+autor:salinas&hl=es&as_sdt=1,5
```

Para ello se ingresa el nombre del autor en el campo *Author Name* y se hace clic en la opción *Search Author* de la solapa *Documents*, el resultado se carga en un objeto *WebBrowser* del cual se obtiene el código HTML resultante (observar Figura 5.24). La opción *Process Record* pasa este contenido a los procesos que se encargan de recorrer todo el contenido mediante un conjunto de expresiones regulares que detectan el inicio y fin de un registro de resultado. Dentro de este ámbito, se detectan las partes del registro de GS como son: Título, Enlace al recurso apuntado, Autores, Año de publicación, *Journal* o *Publication source*, *Abstract*, Cantidad de citas recibidas, y enlace a los documentos que citan. Cuando se especificó el modelo, se establecieron las condiciones que debe reunir un registro para ser incorporado a la base de datos, como la existencia de los Autores y el Año de publicación.

Google Scholar entrega 20 resultados por página, con lo cual se deben recorrer una a una las páginas de resultados desde la misma interfaz mostrada, procesando una a una con la opción *Process Record*.

Luego de procesadas todas las páginas de resultados, en la solapa

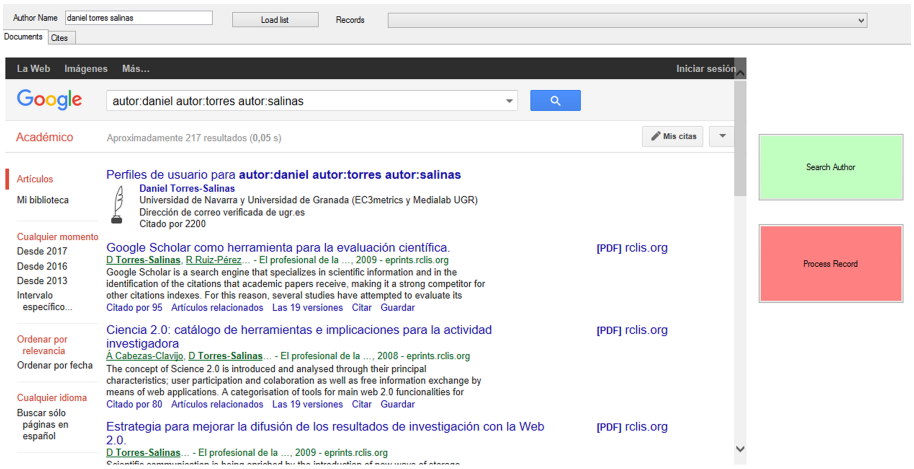


Figura 5.24: Búsqueda por autor en GS desde aplicación de escritorio

Cites se cargan los registros recolectados por el autor desde la opción *Load List* en el control de tipo "lista desplegable". Al elegir un elemento de la lista, en el objeto *WebBrowser* se carga la URL de los documentos que citan a este (ver Figura 5.25).

Este conjunto de registros se recorre nuevamente uno a uno para obtener las páginas de citas, se carga la URL de esta página y se procesan de la misma forma que en el caso anterior pero desde la opción *Process Cites*. Si la página de citas posee más de una página (más de 20 citas) se deben recorrer una a una e ir procesándolas del mismo modo que se hizo con los resultados de publicaciones. Es un proceso largo y no tan automático como se desea, pero es efectivo.

El proceso de recolección finaliza al recolectar el conjunto de publicaciones y citas del autor, luego de esto ya se puede iniciar el procesamiento de los datos.

5.2.2. Procesamiento inicial

Desde este punto en adelante, el procesamiento es idéntico ya sea que se hayan recolectado los registros de GS con el *crawler* descrito, o se hayan recolectado por medio de la AK API o se hayan importado desde una fuente externa. Lo único que hará variar los resultados de los dis-

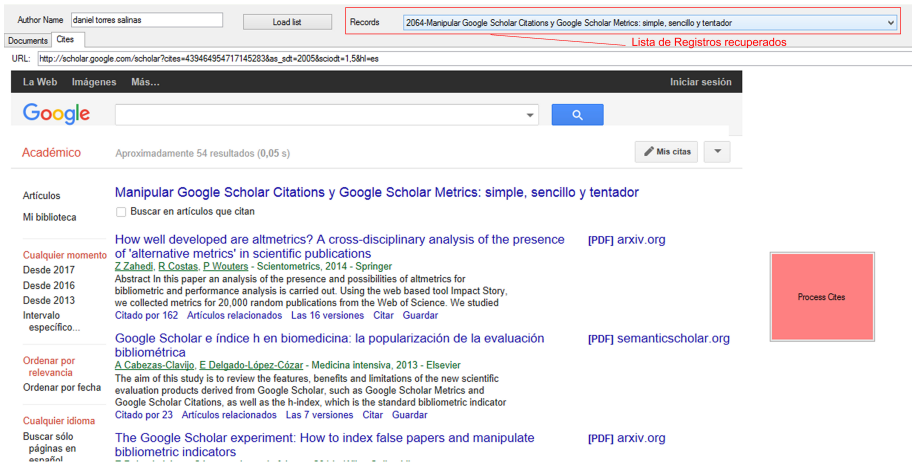


Figura 5.25: Procesamiento de las citas de los registros recolectados desde GS

tintos procedimientos serán la cantidades de agrupaciones y duplicados detectados, y las distintas formas de los nombres de los autores.

Si bien *Microsoft Academic* ofrece cierto grado de control en lo que a nombres de autores se refiere, no es el caso de *Google Scholar*, este motor no siempre indexa el mismo nombre de autor del mismo modo, por ello la gran variabilidad de este campo, además un registro de resultados no posee el listado completo de autores (en caso de que sean mas de dos autores este campo puede aparecer recortado).

Estas pequeñas cuestiones impactan en gran medida en los procesos de desambiguación al aumentar la cantidad de búsquedas, verificaciones, recorridos y cálculos correspondientes. Y no solo eso, *Google Scholar* indexa a veces el nombre completo del autor o partes del mismo, conduciendo esto a la creación de un gran número de agrupaciones de autores.

La Figura 5.26 muestra el resultado del proceso inicial para el autor "Enrique Orduña Malea" indicando como fuente de datos *Google Scholar* (campo *Source* = GS). Si se compara este resultado con el obtenido utilizando como origen de datos a *Microsoft Academic*, las diferencias son más que notables.

Para este experimento se obtuvieron 17 agrupaciones contra las 12

encontradas en el experimento anterior, esto responde primero a la gran variabilidad de los nombres indexados por GS (por ejemplo existen dos variantes para el autor JA Ontalba Ruiperez y JA Ontalba), y segundo, al no aparecer el listado completo de autores de la publicación, esto acarrea otra consecuencia no visible a primera vista, y es que la cantidad de coautores detectados en este experimento, es mucho menor a los encontrados partiendo de MA.

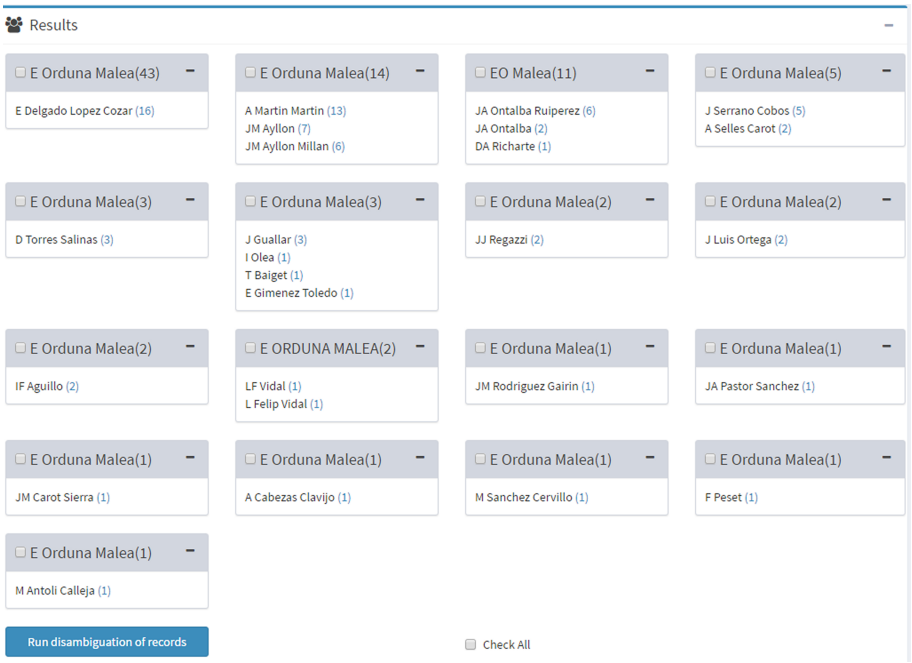


Figura 5.26: Listado de agrupaciones iniciales (GS)

5.2.3. Reproceso

Si hubiese registros a reprocesar, se realizan dos tipos de consulta. En la primera se utiliza la *Academic Knowledge API* para recuperar las publicaciones utilizando el título de la publicación como parámetro de búsqueda. De los resultados obtenidos, por cada uno se revisa el listado de autores para verificar si el autor analizado se encuentra entre estos, si es así, este registro se anexa al listado de publicaciones de la variante de nombre de autor más productivo encontrado hasta el momento (el autor

que posee mayor número de publicaciones).

Si no se encuentra el nombre del autor en ninguno de los resultados obtenidos, es necesario utilizar la *Bing Web Search API* para obtener el top 10 de resultados buscando nuevamente por el título de la publicación. De los resultados obtenidos se recupera la URL al recurso al que apunta y sobre este se realiza una búsqueda mediante expresiones regulares del nombre del autor analizado. Con el primer resultado satisfactorio, el registro se anexa al listado de publicaciones de la variante de nombre de autor más productivo, por el contrario si el nombre no se encuentra en ninguno de los 10 resultados obtenidos, este registro se descarta de forma definitiva.

Para este ejemplo no existen registros a reprocesar, pues se logró establecer la correcta autoría de los registros, es decir, los registros pertenecen al autor o al menos a alguna de sus variantes. No obstante se mostró en el **Capítulo 3** que para una búsqueda del autor "Emilio Delgado López Cózar", *Google Scholar* entrega registros correctos pero mal formados (el nombre completo se encuentra separado como si fuesen dos autores distintos o directamente el nombre del autor no aparece), en estos casos es necesario que se realice el reproceso para establecer si este registro pertenece al autor analizado o debe descartarse.

En otro experimento realizado, sobre las publicaciones recolectadas para el autor Daniel Torres Salinas desde *Google Scholar* se obtuvieron un total de 41 registros que no pudieron ser comprobados, es decir, ni por las combinaciones de los nombres de autor ni buscando cada artículo en la web se pudo comprobar que el autor Daniel Torres Salinas es autor de las mismas, con lo cual estos registros se descartan y no forman parte del conjunto de resultados, en la Tabla 5.3 se muestra el listado de registros en esta situación.

Tabla: 5.3: Registros no comprobables para el autor Daniel Torres Salinas

#	**Título**	**Autores**	**Año**	**Citas**
1	Absceso hepático por Actinomyces. Comunicación de un caso y revisión de la bibliografía 537	, JAM Mendoza, VMV Rodríguez, JG Salinas	2009	2
2	Actas de Diseño Nº 10	, M Roca, H Rondina, AL Russo, R Sacchet, D Salinas	2011	0
3	Actitudes hacia el aprendizaje de las ciencias físicas, naturales y matemáticas de BUP y COU: Un estudio sobre tres dimensiones	, I Fernández, CA Jaime, LES Torres	1993	7
4	Actualidades científicas: 2a. serie de artículos publicados en el Diario Católico Argentino	, EF Basanta, MV Giubergia, JP Martelotto, EE Salinas	1938	0
5	Acute Cardiovascular Care 2014	, H Del Castillo, GL Alonso-Salinas	2014	0
6	Adicciones: un abordaje interdisciplinario	, CO Cervino, JM Affanni, B Salinas, JJB Salinas	2015	0
7	Análisis descriptivo de unidades caprinas en el suroeste de la región lagunera, Coahuila, México	HS González, EDV Moysen	2016	0
8	Análisis y difusión de buenas prácticas agrícolas en el cultivo de frijol mediante la implementación de comunidades de práctica. Iniciativa de innovación tecnológica,	, J Barrera Violeth, C Cardona Ayala, C Salinas	2013	0
9	Aporte de la fijación biológica de nitrógeno a la emisión de NzO desde el suelo con cultivo de soja	, WJ LENNE, JMM MARTINEZ, F REYES, C SALINAS	2001	0
10	Association of a low-frequency variant in HNF1A with type 2 diabetes in a Latino population	, CA Haiman, T Tusié-Luna, CA Aguilar-Salinas	2014	54
11	Compilación de Proyectos de Investigacion de 1984-2002	, J García Montiel, MA Torres Duran, C Torres Márquez	2012	0
12	Compilación de Proyectos de Investigación desde el año 2003 al 2012	, S Fuenlabrada Velázquez, MR Salinas Tobon	2012	0

#	**Título**	**Autores**	**Año**	**Citas**
13	Conocer para prevenir: profundizando en el conocimiento y prevención de la Hepatitis B en la población de estudiantes de la Universidad Católica de Córdoba	, EEJ Cardoso, MCDE SALINAS CARMONA	2013	0
14	Contabilidad de costos	, L Jacobsen, T Salinas, ASAST Salinas	2012	1
15	Contributions of the Cherenkov Telescope Array (CTA) to the 6th International Symposium on High-Energy Gamma-Ray Astronomy (Gamma 2016)	, M Tornikoski, DF Torres, M Torres	2016	0
16	Cuerpos de psamoma y cambios degenerativos de tumores en los plexos coroideos. Estudio clínico-patológico	MLT Suck, DR Bojórquez, CS Lara, RV Orozco	2009	1
17	Cuévano de aventuras	, GV Raul, MCBP Puente Salinas	1935	0
18	Difusión y transferencia de habilidades en el desarrollo de productos de cuero de cabra y jabones para la industria minera en la provincia de Choapa	, ZBM Enrique, ZZN del Carmen, ZMT Tapia, DR Torres	2011	0
19	Efecto de la deficiencia de potasio sobre parámetros foliares de palma de aceite.	JAS Rojas, DGC Salinas	2015	0
20	ESCO (ESCUELA SUPERIOR DE COMUNICACIÓN), CENTRO ADSCRITO A LA UNIVERSIDAD DE GALES, UNIVERSIDAD DE GRANADA Y UNIVERSIDAD	RR CABALLERO, EDLCY DANIEL, T SALINAS	1997	0
21	Estudio crítico de los delitos contra la propiedad: hurto y robo	, D Friz Donoso, M Rojas Sepúlveda, M Salinas Ramos	2005	0
22	Gran cultura confucianaa proposito del 56 aniversario de la Republica Popular China	, RJRJ Soria, ME Guerrero Salinas, MEG Salinas	2006	0
23	Guías de rehabilitación para niños con enfermedades respiratorias crónicas	, R Vera Uribe, R Torres, CY Kuo, P Salinas	2007	25

#	Título	Autores	Año	Citas
24	Immune Response to Nocardia brasiliensisAntigens in an Experimental Model of Actinomycetoma in BALB/c Mice	MC Salinas-Carmona, E Torres-Lopez	1999	62
25	INCORPORACIÓN DE APLICACIONES Y PRÁCTICAS DE COMERCIO ELECTRÓNICO EN MICRO Y PEQUEÑAS EMPRESAS PROVEEDORAS DE LA ZONA	, IC Limitada, IAM Rivera, ILF Balbuena, IED Torres	2011	0
26	Indicadores de ciencia y tecnología Colombia 2013	, I Perea, E Bueno, A Guevara, JM Salinas Pico	2016	2
27	Indicadores de ciencia y tecnología Colombia 2014	, H Mora, G Inés Perea, A Guevara, JM Salinas Pico	2016	0
28	La comunicación como herramienta clave del community manager. Justificación de su presencia en las facultades de comunicación/Communication as a	, MEB Monferrer, MM Camacho, LB Salinas	2012	0
29	Los chinos, sibaritas	, RJRJ Soria, ME Guerrero Salinas, MEG Salinas	2006	0
30	Los religiosos, hoy y mañana	, AT de Biblioni, CA Tapia, R Torres, DDR Torres	1966	0
31	Métodos y mediciones	, R Raimann, FR Chamorro, GRM Torres	2010	0
32	Nocardia brasiliensis Immune Response to	, MC Salinas-Carmona, AI Ernesto Torres-Lopez	1999	0
33	Pastura test: red de ensayos de variedades forrajeras; resultados de la campaña 2013/14.	, SD GALLEGOS, CE IBARRA, IF TORRES	2015	0
34	Percepciones de México: a través de una colección de tarjetas postales	, E Torres, SH Vargas, LMS Franco, CMSE Torres	2014	0

#	Título	Autores	Año	Citas
35	Rare variants in PPARG with decreased activity in adipocyte differentiation are associated with increased risk of type 2 diabetes	, BE Henderson, CAA Salinas	2014	37
36	Salud laboral: prevención de riesgos en el trabajo: aplicación al sector salud en la República Argentina	, MD Arostegui, A Madoz, V Martínez, M JC Salinas	2009	0
37	Serum procollagen type III peptide as a marker of hepatic fibrogenesis in alcoholic hepatitis	M Torres-Salinas, A Parés, J Caballería, W Jiménez	1986	88
38	TRANSFERENCIADE TECNOLOGÍAS EN GESTIÓN DE NEGOCIOS DE LA AGROINDUSTRIAPARA LAAGRUPACIÓN DE 105EMPRESARIOS LA GARZA SA	, WAA Pueyes, J Zambrazambra, DR TORRES	2011	0
39	Tratado de medicina cardiovascular	, D Silvia, C Ciancaglini, RG Lopez Santio, C SALINAS	2008	4
40	Utilidad de los criterios del NCEP y la FID del síndrome metabólico para detectar resistencia a la insulina evaluada por QUICKI	, JC Madrigal, GFH Torres, MZ Juárez, JG Salinas	2009	0
41	Verde que te quiero verde. Educación ambiental en el Jardín Botánico (S0 (BGaspar Xuárez sj (S1 (B de la Universidad Católica de Córdoba	, A da Silva, EC Arevalo, AF Sánchez, EG Torres	1939	0

5.2.4. Análisis de transitividad

El análisis de transitividad, cuyo objetivo es intentar reagrupar los distintos grupos de autores, arrojó como resultado final para este experimento solo 3 agrupaciones (ver Tabla 5.4), un número mucho menor que los obtenidos con los datos de *Microsoft Academic* (7 agrupaciones, ver Figura 5.12), esto responde a que los nombres registrados en MA son mucho más específicos (poseen los nombres y apellidos completos) y están mucho más normalizados con pocas o ningunas variantes en la forma de firmar, en cambio en GS la mayoría de los nombres son almacenados con iniciales, con el orden intercambiado (Apellido y Nombre y/o Nombre y Apellido), con algunos de los nombre y apellidos, en dos partes (Nombre 1 Apellido 1 y Nombre 2 Apellido 2, entendiéndose esto como dos autores diferentes), entre tantas otras formas.

Si bien, puede que parezca un problema que GS indexe los nombres abreviados, separados u omitiendo algún nombre o apellido, esto no es del todo cierto, en los experimentos llevados a cabo, y en las constantes pruebas de los procedimientos, sobre todo del procedimiento *Análisis de transitividad*, se encontró que para obtener mejores resultados es preferible utilizar menos restricciones. Mientras más específico sea un nombre de autor (contenga todos los apellidos y todos los nombres) los resultados se verán más limitados, mientras que si se utilizan iniciales o abreviaciones u omisiones en algunos de los apellidos y/o nombres (como es el caso de los autores indexados por GS), los resultados mejoran notablemente al ampliar el rango de búsqueda y reducir en cierta forma las restricciones.

Tabla: 5.4: Listado de agrupaciones final para GS

#	**Nombre de autor**	**Coautores**
1	E Orduna Malea	E Delgado Lopez Cozar, A Martin Martin, JM Ayllon, JM Ayllon Millan, JA Ontalba Ruiperez, J Serrano Cobos, J Guallar, D Torres Salinas, JJ Regazzi, A Selles Carot, J Luis Ortega, JA Ontalba, IF Aguillo, M Antoli Calleja, DA Richarte, JA Pastor Sanchez, A Cabezas Clavijo, F Peset, T Baiget, E Gimenez, Toledo, LF Vidal, L Felip Vidal, JM Rodriguez Gairin, I Olea
2	E Orduna Malea	JM Carot Sierra

3	E Orduna Malea	M Sanchez Cervillo

5.2.5. Detección de duplicados, desambiguación de registros

Un resultado interesante y no tan esperado, es la reducida cantidad de elementos duplicados detectados con estos datos frente a los encontrados utilizando los datos de *Microsoft Academic*. Existen registros duplicados, pero en menor medida, al menos para los autores y para los conjuntos de datos recolectados y analizados. No es una afirmación absoluta, pero se puede ver que *Google Scholar* ha invertido esfuerzos en reducir estos desfasajes.

El resultado de la detección de registros duplicados se observa en la Tabla 5.5.

Tabla: 5.5: Listado de registros duplicados resultante del análisis (GS)

#	**Título**	**Año**	**Citas**
1	Espacio universitario español en la Web (2010): estudio descriptivo de instituciones y productos académicos a través del análisis de subdominios y subdirectorios	2013	5
2	Espacio universitario español en la Web (2010): estudio descriptivo de instituciones y productos académicos a través del análisis de subdominios y subdirectorios	2013	2
3	Impacto de los repositorios a través de técnicas cibermétricas: el caso general de Latinoamérica y especial de Costa Rica	2013	1
4	Impacto de los repositorios a través de técnicas cibermétricas: el caso general de Latinoamérica y especial de Costa Rica	2013	0

Las filas de color azul indican al registro que se seleccionó como principal, mientras que el o los registros debajo de color blanco son los duplicados.

En la fila número 1 se puede observar que el registros posee 5 citas, mientras que el duplicado (fila número 2) posee 2 citas, en estos casos se realiza la fusión de citas. Este procedimiento es posible en este experimento dado que se pudieron recuperar de GS las publicaciones que citan.

El registro de la fila 1 posee el identificador de registro 2444 mientras que el registro de la fila 2 posee el identificador 2456 (estos identificadores son las claves principales de la tabla *Records* donde estos registros fueron almacenados). La Tabla 5.6 lista los registros de citas para ambas publicaciones, como se observa, las filas 2 y 7 correspondientes a las registros 2444 y 2456 respectivamente, son duplicados, con lo cual, el número real de citas de la publicación en cuestión no son 5 ni 2 ni 7, sino 6 citas, este es un claro ejemplo de la fusión de citas.

Tabla: 5.6: Listado de citas de los registros duplicados

#	Id	Título	Año
1	2444	Acceso al conocimiento público universitario en España: patrones geográficos	2015
2	2444	Are web mentions accurate substitutes for inlinks for Spanish universities?	2014
3	2444	CALIMACO: desarrollo de un servicio de bibliotecario virtual para la interacción multimodal con dispositivos móviles	2016
4	2444	From universities to private companies: a measurable route of LinkedIn users	2016
5	2444	Visibilidad e impacto web de los grupos de investigación de información y documentación en las universidades públicas españolas	2013
6	2456	Aggregation of the web performance of internal university units as a method of quantitative analysis of a university system: The case of Spain	2013
7	2456	Are web mentions accurate substitutes for inlinks for Spanish universities?	2014

Una vez finalizado el proceso de desambiguación de registros, la aplicación ofrece un panel de resultados como el mostrado en la Figura 5.27.

En la Figura 5.28 se ofrece a modo comparativo, los indicadores obtenidos con *Academic Evaluator* para ambos experimentos. Utilizando como orígenes de datos a *Google Scholar* y *Microsoft Academic*.

Es notable la diferencia en todos los indicadores, es esperable que al aumentar la cantidad de documentos aumentase también la cantidad de citas recibidas, pero comparando estos orígenes de datos tan distintos no es así. GS con 92 documentos recoge 289 citas, mientras que MA con 107 solo es capaz de recoger 185 citas.

Figura 5.27: Panel de resultados para el autor Enrique Orduña Malea, *source*=GS

Claramente la reducción en el número de citas afectará los indicadores que dependen de éstas (*h*-index, *g*-index, *hg*-index e *i10*-index). Otro aspecto a destacar tiene que ver con la cantidad de documentos duplicados para cada origen de datos, MA posee la mayor cantidad de registros recuperados y duplicados.

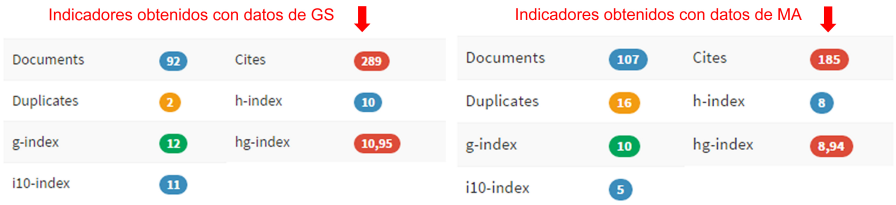


Figura 5.28: Comparación de indicadores obtenidos en AE con los datos de GS y MA

5.2.6. Visualización de resultados

Las opciones de visualización *View chart before* y *View chart after*, ofrecen las siguientes gráficas (ver Figura 5.29 y Figura 5.30 respectivamente)

Al existir tan pocos elementos duplicados, las gráficas son muy similares salvo el pequeño detalle al que se hace mención a propósito posando el mouse, la cantidad de citas fusionada para el registro duplicado de *"Espacio universitario español en la Web (2010): estudio descriptivo de instituciones y productos académicos a través del análisis de subdominios y subdirectorios"*.

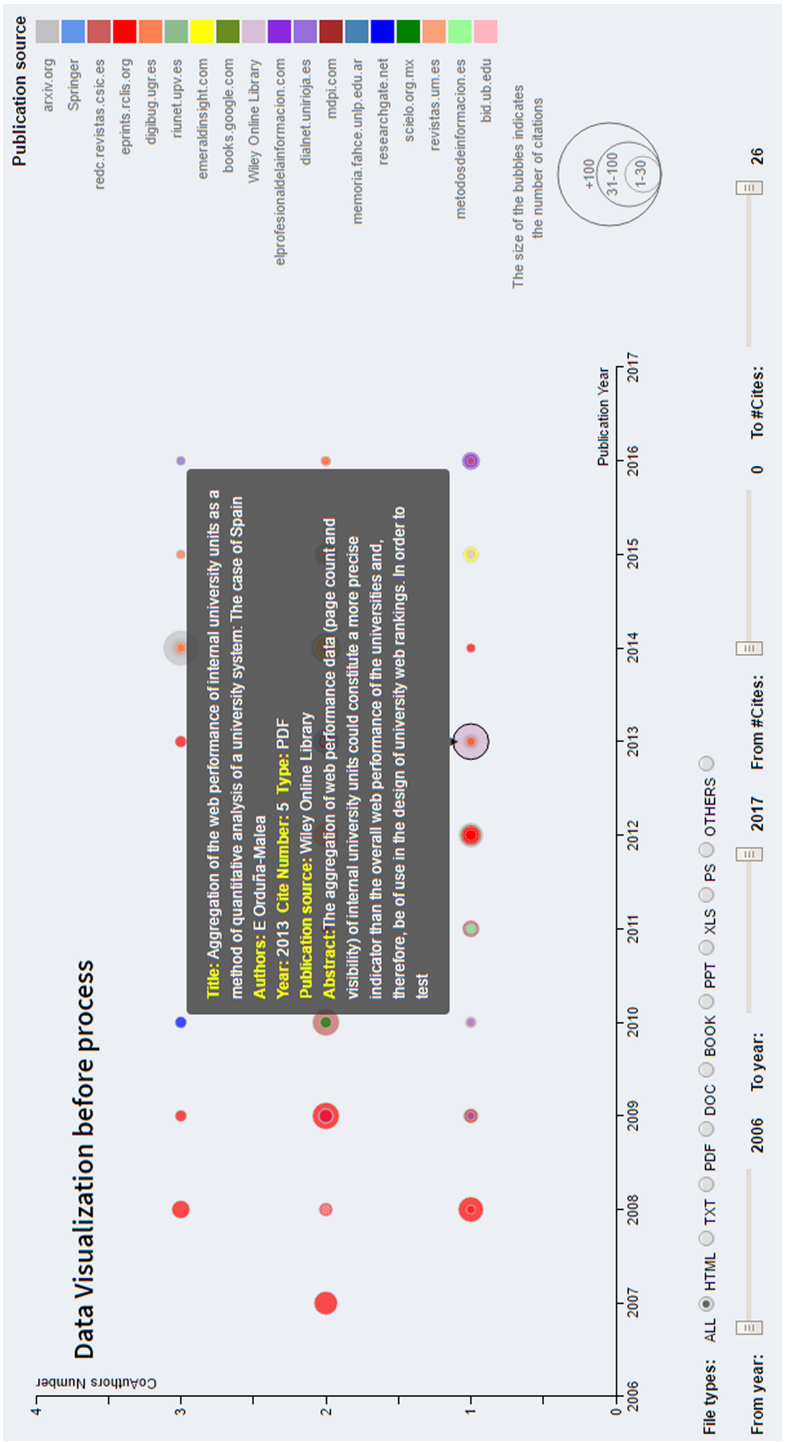


Figura 5.29: Visualización de *scatterplot* antes del análisis (GS)

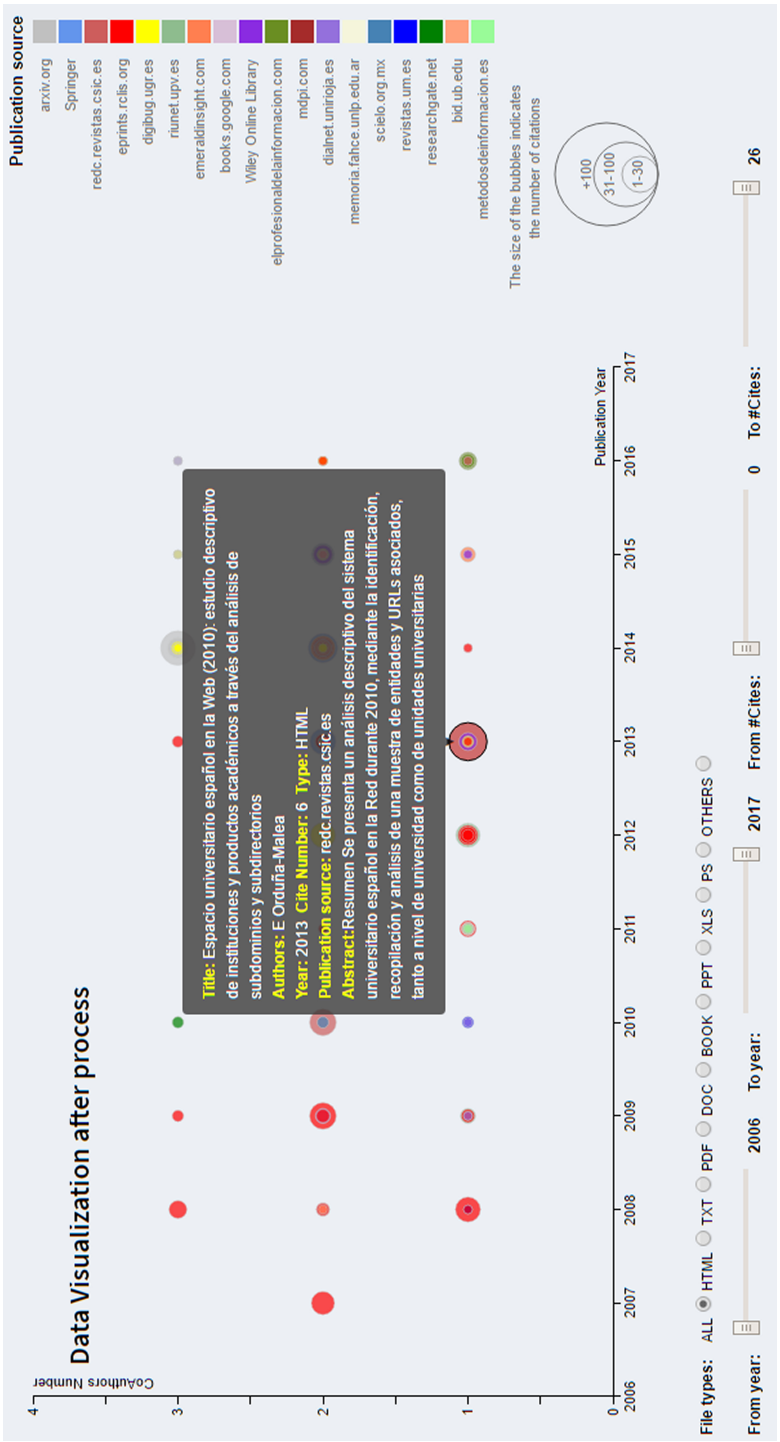


Figura 5.30: Visualización de *scatterplot* luego del análisis (GS)

5.3. Análisis de resultados

El modelo aquí planteado y por consiguiente la aplicación que da soporte al mismo (*Academic Evaluator*) ofrecen ciertas ventajas frente a otro tipo de herramientas de evaluación científica. Las ventajas radican en la flexibilidad que aporta al utilizar cualquier origen de datos, también en el modo en que selecciona las agrupaciones resultantes (dejando a elección del usuario la decisión final de agruparlas o no) y sobre todo en la desambiguación de publicaciones y detección de duplicados. Este último ítem es algo frecuente que se le ha reclamado en muchísimos trabajos a *Google Scholar* y seguramente no tardarán en reclamárselo a *Microsoft Academic*.

Las herramientas y la gran parte de los estudios, análisis e investigaciones que utilizan como origen de datos a *Google Scholar*, se limitan a informar de la existencia de duplicados pero no se encargan de resolver el problema de fondo, otras en cambio simplemente hacen caso omiso a ello, caso del tan aclamado software *Publish or Perish*.

La idea de este apartado es simplemente ofrecer un resumen de los resultados obtenidos en los experimentos realizados empleando la herramienta *Academic Evaluator* tomando como orígenes de datos a MA y GS, frente a los resultados obtenidos por las herramientas y bases de datos tradicionales. Oportunamente, se hizo mención del estado actual de las herramientas más importantes y las más utilizadas para llevar a cabo análisis bibliométricos en lo que respecta a indicadores, opciones de gráficos estadísticos y visualizaciones de información. La Tabla 5.7 muestra un resumen de los resultados entregados por:

1. *Google Scholar*
2. *Google Scholar Citations*
3. *Microsoft Academic*
4. *Microsoft Academic v2 Preview*
5. *Academic Evaluator* utilizando como origen de datos a *Microsoft Academic*
6. *Academic Evaluator* utilizando como origen de datos a *Google Scholar*

7. *Publish or Perish* utilizando como origen de datos a *Microsoft Academic*
8. *Publish or Perish* utilizando como origen de datos a *Google Scholar*
9. *Web of Science*
10. *Scopus*

Donde el detalle de las columnas es el siguiente:

- *Documentos*: muestra el total de los documentos con los que se inicia el análisis o el total de documentos recolectados.
- *Documentos finales*: muestra el total de documentos luego de filtrados, procesados y desambiguados el conjunto inicial, como la única herramienta que realiza estas acciones es *Academic Evaluator*, es la única en donde los valores de esta columna difieren de la columna *Documentos*.
- *Citas*: muestra el total de citas recibidas por el conjunto de publicaciones.
- h-index: valor del h-index
- g-index: valor del g-index
- hg-index: valor del hg-index
- $i10$-index: valor del $i10$-index
- *Coautores*: cantidad de coautores resultantes de las publicaciones analizadas
- *Duplicados*: cantidad de duplicados detectados. Nuevamente, la única herramienta que realiza detección de duplicados es *Academic Evaluator*, sin embargo se indicó la cantidad de registros duplicados encontrados de forma manual en *Google Scholar* (2 duplicados), *Publish or Perish* basado en *Microsoft Academic* (16 duplicados) y *Publish or Perish* basado en *Google Scholar* (8 duplicados).

Tabla: 5.7: Indicadores entregados por las herramientas bibliométricas más utilizadas

#	**Herramienta**	**Documentos**	**Documentos finales**	**Citas**	**h-index**	**g-index**	**hg-index**	**i10-index**	**Coautores**	**Duplicados**
1	Google Scholar	96	96							2
2	Google Scholar Citations	115	115	492	13			16		
3	Microsoft Academic	135	135							
4	Microsoft Academic v2 Preview	102	102	182						
5	AEvaluator basado en *Microsoft Academic*	130	113	188	8	10	8,94	5	52	16*
6	AEvaluator basado en *Google Scholar*	94	92	285	10	12	10,95	11	26	2
7	PoP basado en *Microsoft Academic*	132	132	190	8	10				16*
8	PoP basado en *Google Scholar*	114	114	450	12	17				8
9	Web of Science	36	36	91	5				27	
10	Scopus	37	37	124	6				28	

Es sabido que a partir de un conjunto de datos, se pueden calcular diversos indicadores y no solo los ofrecidos por estas herramientas, en la tabla presentada se muestran los valores de los indicadores que entrega cada herramienta y no los que se podrían calcular con cada una de ellas. En todas las herramientas se realizaron búsquedas por "nombre de autor" para el autor "Enrique Orduña Malea".

A simple vista es evidente la cobertura de las distintas herramientas observando la tercera columna (Documentos), las menos favorecidas son *Scopus* y *Web of Science*, con 37 y 36 registros respectivamente, frente *Microsoft Academic* (135 registros) o las herramientas que utilizan a este motor como origen de datos: PoP (132 registros) y AE (130 registros).

La relación que existe entre la cantidad de publicaciones y el número de citas se puede observar en la Figura 5.31. La misma hace evidente la superioridad numérica en cuanto a las citas que recoge *Google Scholar* frente a las que puede recolectar *Microsoft Academic*. Si bien, al parecer MA ha recuperado un mayor número de documentos que GS, es MA quien posee la mayor cantidad de registros duplicados.

En cuanto a los indicadores ofrecidos, el que casi es un denominador común es el h-index, y es una pena que al realizar esta evaluación, las bases de datos tradicionales como *Scopus* y *WoS* no presenten un punto de comparación, pues solo recolectan menos del 30 % de los registros que recolectan motores de libre acceso.

Un dato que vale la pena aclarar, la recolección de registros de *Google Scholar* para ser procesados por *Academic Evaluator* (94 registros iniciales), fueron recolectados el día 12-12-2016, por ese motivo no están incluidos los registros con año de publicación 2017 (que son dos documentos y que suman un total de 96 como se indica en la fila correspondiente a GS).

En cuanto a la desambiguación y detección de duplicados, ya se aclaró que solo *Academic Evaluator* es capaz de detectar duplicados y descartar estos registros de los resultados finales y cálculo de indicadores, no así una de las herramientas muy conocidas por utilizar como origen de datos a *Google Scholar* y que recientemente incorporó a *Microsoft Academic*, *Publish or Perish*.

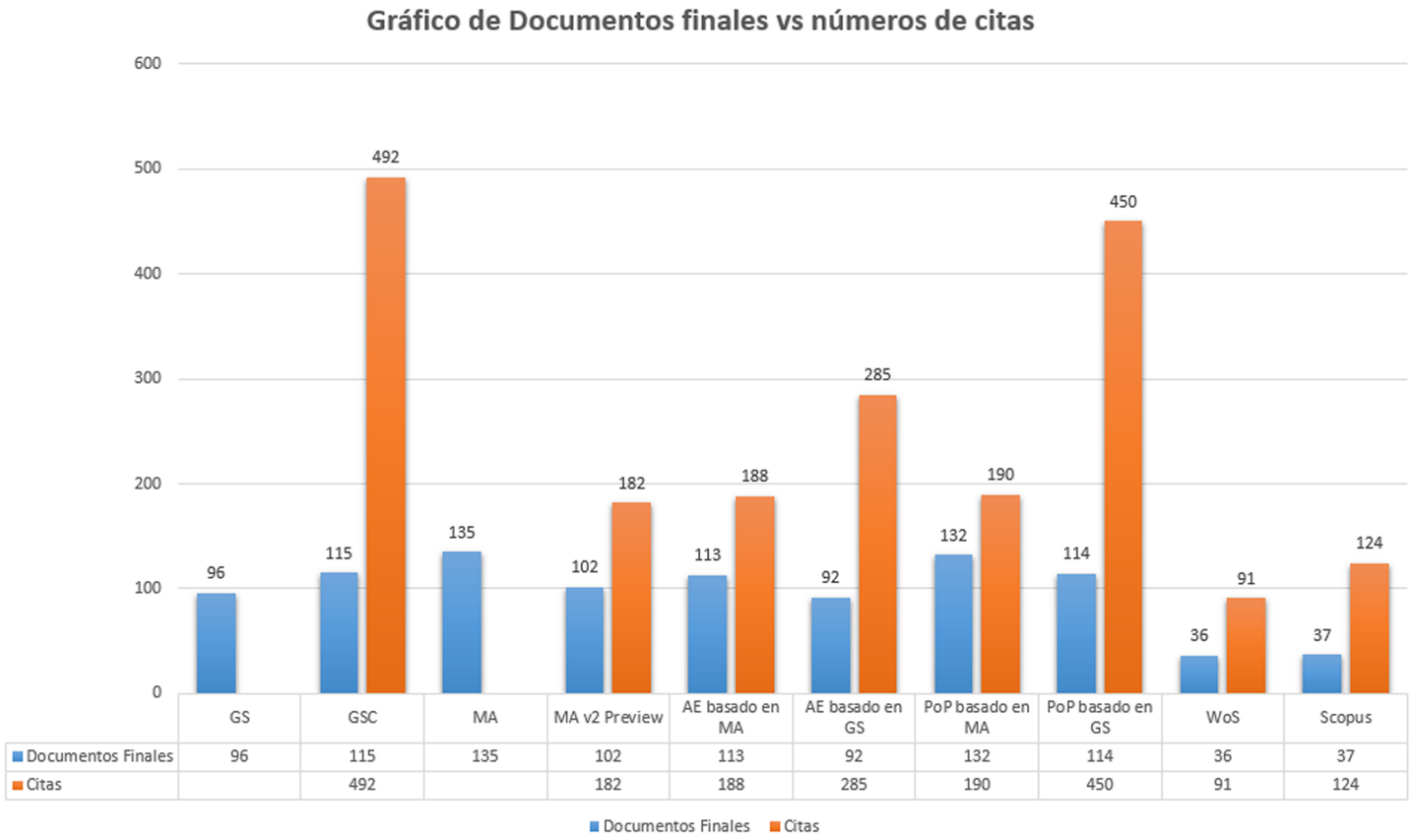


Figura 5.31: Gráfico de barras de Documentos finales vs número de citas

PoP es una herramienta muy potente, versátil, intuitiva, con un gran número de opciones y parámetros de búsqueda, filtrado, y posibilidades de exportar los datos, además de que es muy utilizada por la comunidad científico-académica, sin embargo le falta un pequeño giro para dar un gran salto cualitativo, y es la posibilidad de detectar los registros duplicados.

Al momento de realizar este análisis comparativo, y como se ve en la tabla 5.7, se realizó una inspección manual y se detectaron que *Publish or Perish* utilizando como origen de datos a *Microsoft Academic* posee 15 registros duplicados y uno triplicado (ver Figura 5.32) y *Publish or Perish* utilizando como origen de datos a *Google Scholar* cuenta con 8 registros duplicados (ver Figura 5.33).

La fila 5 de la Tabla 5.7, muestra los resultados para *Academic Evaluator* basado en *Microsoft Academic*, se parte de un conjunto de 130 registros y el resultado final son 113 documentos con 16 duplicados, el resultado final es 129 (113 + 16), el registro faltante se debe a que, de los 16 registros duplicados, 15 son realmente duplicados y solo 1 triplicado.

Otro dato a destacar en esta comparación, es la cantidad de coautores detectados por *Academic Evaluator*. Al utilizar como origen de datos de MA encontró 52 coautores y al utilizar GS encontró solo 26. Esto se debe principalmente a que *Microsoft Academic* almacena el conjunto completo de autores de una publicación y los facilita a través de la AK API.

En cambio *Google Scholar* indexa los autores que firman una publicación pero no ofrece el conjunto completo de autores debido a que posee un espacio reducido para esta información. Diferente es el caso de *Google Scholar Citations*, donde es el propio autor quien gestiona las publicaciones y puede agregar el listado completo de los autores de dichas publicaciones.

Una alternativa más que interesante sería poder realizar la recolección de publicaciones desde el perfil del autor en GSC, pero a día de hoy no todos los científicos poseen un perfil creado, razón por la cual la recolección se sigue realizando desde el panel principal de *Google Scholar*.

Cites	Per year	Rank	Authors	Title	Year	Publication
3	0.75	24	Enrique Ordunam...	Aggregation of the web performance of internal university units as a method of quantitative analysis of a university system: The case of Spain	2013	Journal of the Association...
1	0.25	38	Enrique Orduna M...	Aggregation of the web performance of internal university units as a method of quantitative analysis of a university system: the case of Spain	2013	Journal of the Association...
0	0.00	81	Enrique Ordunam...	Aggregation of the web performance of internal university units as a method of quantitative analysis of a university system: The case of Spain...	2013	Journal of the Association...
0	0.00	109	Enrique Ordunam...	Análisis de la correlación entre la audiencia web de los medios digitales de prensa española y su visibilidad en gestores sociales en noticias	2010	
0	0.00	127	Enrique Orduna M...	Análisis de la correlación entre la audiencia web de los medios digitales de prensa española y su visibilidad en gestores sociales en noticias	2010	
0	0.00	131	Enrique Orduna M...	Análisis de los FRBR en la ejecución de tareas genéricas en catálogos compartidos	2008	
0	0.00	132	Enrique Ordunam...	Análisis de los FRBR en la ejecución de tareas genéricas en catálogos compartidos	2008	
0	0.00	116	Enrique Orduna M...	"Attention profile": ¿información a la que prestamos atención?	2009	
0	0.00	107	Enrique Ordunam...	Attention profile: ¿información a la que prestamos atención?	2009	
0	0.00	110	Enrique Ordunam...	El proyecto visado arquitectnico: descripcin, caracterizacin y normalizacin documental	2009	Profesional De La Informa...
0	0.00	90	Enrique Ordunam...	El proyecto visado arquitectónico: descripción, caracterización y normalización documental	2009	Profesional De La Informa...
0	0.00	123	Enrique Orduna M...	Fuentes de enlaces web para análisis cibermétricos	2012	
1	0.20	40	Enrique Ordunam...	Fuentes de enlaces web para análisis cibermétricos (2012)	2012	
3	0.60	21	Enrique Ordunam...	Graphic, multimedia, and blog content presence in the Spanish academic web-space	2012	
1	0.20	41	Enrique Orduna M...	Graphic, multimedia, and blog content presence in the Spanish academic web-space	2012	
4	1.33	17	Emilio Delgadolop...	H Index Scholar: el índice h de los profesores de las universidades públicas españolas en humanidades y ciencias sociales	2014	Profesional De La Informa...
0	0.00	49	Emilio Delgado Lo...	H-Index Scholar: el índice h de los profesores de las universidades públicas españolas en humanidades y ciencias sociales	2014	Profesional De La Informa...
0	0.00	120	Enrique Ordunam...	Impacto de los repositorios a través de técnicas cibermétricas: el caso general de Latinoamérica y especial de Costa Rica	2013	
0	0.00	122	Enrique Ordunam...	Impacto de los repositorios a través de técnicas cibermétricas: el caso general de Latinoamérica y especial de Costa Rica [material complemen...	2013	
0	0.00	72	Enrique Ordunam...	La bibliometría que viene: ALMetrics (Author Level Metrics) y las múltiples caras del impacto de un autor	2016	Profesional De La Informa...
0	0.00	74	Enrique Orduna M...	La bibliometría que viene: ALMetrics (Author Level Metrics) y las múltiples caras del impacto de un autor	2016	Profesional De La Informa...
0	0.00	105	Enrique Orduna M...	Personalización e interactividad en los rankings de universiades publicados en la Web	2011	
4	0.67	15	Enrique Ordunam...	Personalización e interactividad en los rankings de universidades publicados en la Web	2011	
1	0.14	30	Alicia Selles Carot,...	Proposal of a Goal-Oriented Shared Catalog Model	2010	International Conference ...
0	0.00	64	Alicia Selles Carot,...	Proposal of a Goal-Oriented Shared Catalog Model	2010	IEEE Internet Computing
1	0.13	31	Enrique Orduna M...	Propuesta de indicadores métricos para gestores sociales de noticias: análisis de la prensa digital española en Menéame	2009	Information Research
0	0.00	106	Enrique Ordunam...	Propuesta de indicadores métricos para gestores sociales de noticias: análisis de la prensa digital española en Menéame	2009	Information Research
0	0.00	108	Enrique Ordunam...	Ranking de universidades en la Unión Europea: aproximación multidimensional a una realidad compleja	2010	
0	0.00	126	Enrique Orduna M...	Ranking de universidades en la Unión Europea: aproximación multidimensional a una realidad compleja	2010	
0	0.00	63	David Azorinrichar...	Redes de conectividad entre empresas tecnológicas a través de un análisis métrico longitudinal de menciones de usuario en Twitter	2016	Revista Espanola De Docu...
0	0.00	113	David Azorin Rich...	Redes de conectividad entre empresas tecnológicas a través de un análisis métrico longitudinal de menciones de usuario en Twitter	2016	Revista Espanola De Docu...
0	0.00	111	Enrique Ordunam...	Reutilización e intercambio de objetos digitales compuestos en la Web: el proyecto OAI-ORE	2009	
0	0.00	117	Enrique Orduna M...	Reutilización e intercambio de objetos digitales compuestos en la Web: el proyecto OAI-ORE	2009	

Figura 5.32: Registros duplicados encontrados en PoP utilizando *Microsoft Academic* como origen de datos

Cites	Per year	Rank	Authors	Title	Year	Publication	Publisher
0	0.00	113	EO Malea, MFP M...	Análisis de la variabilidad de nombres de autor españoles en depósitos digitales universitarios de acceso abierto: un estudio por áreas de...	2009	Revista española de ...	riunet.upv.es
2	0.25	61	E Orduña-Malea, ...	Análisis de la variabilidad de nombres de autores españoles en depósitos digitales universitarios de acceso abierto: un estudio por áreas ...	2009	Revista ...	Consejo Superior de ...
0	0.00	97	EO Malea	Attention profile¿ información a la que prestamos atención?	2009	Anuario ThinkEPI	riunet.upv.es
0	0.00	98	E Orduña-Malea	Attention profile¿ información a la que prestamos atención?	2008	Anuario ThinkEPI	EPI SCP, Barcelona, Sp
1	0.13	66	E Orduña-Malea, J...	El Grupo ThinkEPI: un think tank en información y documentación	2009		Fesabid
0	0.00	85	EO Malea, J Gualla...	El Grupo ThinkEPI: un think tank en información y documentación	2009	... Españolas de Documen...	Fesabid
1	0.25	68	A Sellés Carot, E O...	Estrategias y oportunidades tecnológicas en la generación de linked data en las bibliotecas	2013	Mi biblioteca	eprints.rclis.org
0	0.00	84	AS Carot, E Orduñ...	Estrategias y oportunidades tecnológicas en la generación de linked data en las bibliotecas	2013		mysciencework.com
0	0.00	86	EO Malea	Fuentes de enlaces web para análisis cibermétricos	2012	Anuario ThinkEPI	riunet.upv.es
7	1.40	29	E Orduña-Malea	Fuentes de enlaces web para análisis cibermétricos (2012)	2012	Anuario ThinkEPI	eprints.rclis.org
1	0.25	71	E Orduña-Malea	Impacto de los repositorios a través de técnicas cibermétricas: el caso general de Latinoamérica y especial de Costa Rica	2013		digibug.ugr.es
0	0.00	99	E Orduña-Malea	Impacto de los repositorios a través de técnicas cibermétricas: el caso general de Latinoamérica y especial de Costa Rica [material compl...	2013		digibug.ugr.es
1	1.00	37	A Martín-Martín, J...	Proceedings Scholar Metrics: H Index of proceedings on Computer Science, Electrical & Electronic Engineering, and Communications a...	2016	arXiv preprint arXiv: ...	arxiv.org
0	0.00	59	A Martín-Martín, J...	Proceedings Scholar Metrics: H Index of proceedings on Computer Science, Electrical & Electronic Engineering, and Communications a...	2016		digibug.ugr.es
0	0.00	112	LF Vidal, E Orduña...	Resum [Resumen][Abstract]		bid.ub.edu	
0	0.00	95	LF Vidal, B de doc...	Resumen [Resum][Abstract]		bid.ub.edu	

Figura 5.33: Registros duplicados encontrados en PoP utilizando *Google Scholar* como origen de datos

Capítulo 6

Conclusiones y futuras líneas de investigación

En este capítulo se describen las conclusiones obtenidas en el trabajo de tesis doctoral. En primer lugar se hace una división de las conclusiones por cada objetivo planteado, luego se resumen las conclusiones generales, seguido de esto se plantean las futuras líneas de investigación y por último se ofrece una análisis FODA del modelo presentado.

6.1. Conclusiones

Al iniciar esta tesis se plantearon un conjunto de objetivos para guiar el desarrollo de la misma, por cada objetivo se ofrecerán las conclusiones obtenidas a lo largo de este proceso. Estos objetivos, conjuntamente, se plantearon para resolver el problema de estimar la producción académica evaluando el conjunto de publicaciones de un autor provenientes de un motor de libre acceso.

Si bien no era un objetivo planteado desde el inicio del proyecto, se desarrolló un prototipo software totalmente funcional llamado *Academic Evaluator*, el cual reúne todos los procesos involucrados en una única aplicación que guía y brinda soporte, desde la obtención de los datos a evaluar hasta la visualización de los resultados.

Objetivo 1: *El primero de ellos es crear una herramienta que permi-*

ta reunir los resultados de dos motores académicos de libre acceso: Google Scholar y/o Microsoft Academic. Para ello se creará un crawler que sea capaz de consultar y procesar en tiempo real las consultas realizadas por el usuario contra estos motores desde una única interfaz.

1. En primera instancia se logró construir un *crawler* para recuperar los registros de publicaciones de *Google Scholar*, se intentó adaptar la recogida automática a la herramienta desarrollada, pero los bloqueos y limitaciones impuestos por *Google* impidieron el normal funcionamiento. Por esta razón, se adaptó dicho procedimiento por separado en una aplicación externa a la desarrollada para realizar la recolección desde este origen de datos.

2. Luego de construido y probado el *crawler* de GS, se construyó el *crawler* para recuperar los registros que provienen de *Microsoft Academic*, esta fue una tarea más sencilla, puesto que *Microsoft* ofrece una API para tal tarea (AK API). Si bien esta API posee ciertas limitaciones, para las consultas utilizadas en este proyecto no hubo ningún inconveniente. Solo como acotación, en medio de las pruebas, *Microsoft* cambió la definición de los metadatos de las APIs, lo cual llevó a readaptar y redefinir los procedimientos de consulta y recuperación.

3. Se logró incorporar la recuperación de los registros de *Microsft Academic* en línea y está totalmente integrada en la herramienta desarrollada para realizar la recolección de forma automática.

4. El mismo mecanismo utilizado para la recolección de los registros de MA fue empleado para realizar consultas por título de publicación y por palabras claves para los procedimientos de desambiguación.

5. Además de construir los *crawlers* correspondientes, se amplió el funcionamiento del proceso de recolección de información para incorporar datos de fuentes externas u otros motores o bases de datos académicas, y poder de este modo contar con múltiples conjuntos de datos de diversas fuentes.

Objetivo 2: *El segundo objetivo es resolver el problema de la ambigüedad de los nombres de autores y de los títulos de las publicaciones. Como se dijo anteriormente, este problema se presenta en la mayoría*

de las fuentes de datos y es una cuestión muy importante a resolver a la hora de determinar la productividad de un determinado autor, grupo o institución científica. En la bibliografía se pueden encontrar diversos trabajos abordando el tema desde distintos puntos de vistas y alternativas para dar solución a los problemas particulares que se estudian, es decir, hoy en día no existe una solución absoluta o genérica para las diferentes variantes de este problema, por ello en este trabajo se estudiarán las distintas variantes y se implementará aquella que mejor se adapte y que entregue un buen balance entre tiempo de ejecución y efectividad. Es necesario destacar que en este trabajo el tiempo de proceso/respuesta es muy importante ya que los resultados y el procesamiento se obtienen en tiempo real.

1. Para resolver todos los problemas derivados de la falta de normalización (imposibilidad de identificar inequívocamente a un autor, imposibilidad de identificar homónimos, registros duplicados, conteo de citas erróneo), se diseñó un proceso novedoso basado en heurísticas que ataca cada variante del problema en dos grandes procesos. El primero resuelve el problema de la ambigüedad de los nombres de autores y el segundo resuelve el problema de la ambigüedad en los títulos de las publicaciones.

2. Para resolver el problema de la ambigüedad de los nombres de autores se separó el proceso en tres etapas, cada etapa corresponde a un conjunto de condiciones específicas:

    - En la primera etapa, a partir del conjunto de publicaciones recolectadas para un autor, se verifica si efectivamente estas publicaciones pertenecen al autor objeto de análisis. Una vez comprobado esto, el procedimiento realiza un análisis de coautoría para asignar a cada autor-coautor la publicación analizada e identificar agrupaciones de coautores basadas en patrones de colaboración, esto logra identificar los homónimos por un lado y por otro lado los distintos grupos de coautores que colaboran entre sí, basándose en el supuesto de que es poco probable que teniendo un autor homónimo, los coautores de ambos grupos colaboren. Para ello se implementó, de forma satisfactoria y con resultados notables, un mecanismo apoyado en reglas de comparaciones de patrones, expresiones regulares y funciones de distancia de edición para detectar to-

das las posibles combinaciones que puede asumir un nombre de autor, y poder así asignar cada autor a un grupo específico.

- La segunda parte del proceso permite establecer la correcta autoría de las publicaciones que no pudieron evaluarse en la primera etapa debido a errores en la indexación de los autores de la misma. El proceso implementado, logró establecer la autoría en todos los casos positivos y en los casos en que no pudo establecerla, se verificó manualmente que dicha publicación no pertenecía al conjunto de publicaciones del autor.
- La tercera etapa logra reagrupar los grupos identificados en la primera etapa, mediante la extracción de conocimiento externo a través de diversas consultas a la web, analizando patrones y expresiones regulares, para establecer o encontrar relaciones entre grupos de autores que no se pudieron establecer con los datos iniciales. Este proceso reagrupó de forma apropiada la mayor parte de los grupos, y en los casos en que no pudo reagruparlos, se verificó que efectivamente no existía relación entre los grupos evaluados, para el caso de autores que poseen más de un grupo de coautores.

3. Una vez terminado el primer proceso, para las agrupaciones de autores resultantes se resolvió el problema de la ambigüedad de los títulos de las publicaciones asignadas a cada autor, para logra esto se implementó un esquema de agrupación de publicaciones por año de publicación y sobre estos grupos se identificaron los posibles duplicados ajustando los umbrales de comparación de distintas funciones de distancia de edición. Con los umbrales correctos, se logró determinar la duplicidad de registros eficazmente. Y una vez identificados los registros duplicados, el proceso realizó la fusión de citas, evaluando nuevamente la existencia de duplicados en el conjunto publicaciones que citan y descartando los mismos, de esta forma, el número de citas de una publicación es un número correcto y libre de los sesgos de la duplicidad.

4. Todos los resultados obtenidos, cálculo de indicadores y gráficos se reflejan en un panel general de resultados, diseñado específicamente para resumir el proceso completo.

Objetivo 3: *Por último, el objetivo final es diseñar una visualización novedosa, ágil e interactiva para el usuario, que involucre la mayor canti-*

dad de dimensiones posibles de los datos extraídos y procesados (autores, número de publicaciones, número de citas). En este sentido se dejarán de lado los simples gráficos estadísticos presentados hasta el momento.

1. Se logró diseñar e implementar una visualización totalmente nueva, con interesantes opciones de interacción y animaciones, que brindan al usuario una experiencia agradable y que conjuga en un mismo gráfico gran parte de las variables y dimensiones del problema en cuestión.

2. La visualización construida permite representar cada una de las publicaciones involucradas, la cantidad de citas recibidas, el año de publicación, la cantidad de coautores, el lugar de publicación o el nombre de la revista, permite obtener un resumen del contenido de la publicación a simple vista, y además permite acceder al recurso enlazado.

3. Proporciona una imagen/fotografía del conjunto total de publicaciones recolectadas de un autor tanto antes del proceso de análisis como después de finalizado, esto ofrece otro parámetro de comparación y evaluación del proceso empleado.

4. Además de las propiedades y ventajas enumeradas, la visualización se adapta a cualquier conjunto de datos que posea el formato correcto, por otro lado, al estar construida con librerías estándares puede ser incorporada como herramienta de visualización de cualquier motor o base de datos académica.

A modo de conclusión general, se comprueba que el modelo de evaluación propuesto entrega buenos resultados y puede presentar no solo una alternativa viable sino una solución más que aproximada a los problemas que padecen los motores de libre acceso. Aquí se ha presentado un modelo que no se limita a exponer el problema y detectar los inconvenientes, sino que resuelve de forma completa las distintas variantes del problema de la ambigüedad derivadas de la utilización de datos poco normalizados. Mediante la utilización de un conjunto de reglas lógicas y mecanismos de inferencia, completa el circuito desde la toma del dato, hasta el cálculo del resultado final. Solo resta preguntarse el alcance que puede obtener este tipo de soluciones.

6.2. Futuras líneas de investigación

Finalmente, se plantean una serie de futuras líneas de investigación que deben ayudar a completar los resultados obtenidos en este trabajo:

- Una alternativa que se puede utilizar para mejorar las tareas de desambiguación, podría ser utilizar un conjunto de palabras que co-ocurran en dos o más documentos donde se asegura que estos pertenecen al autor objeto de estudio, de esta forma documentos donde no se tiene la certeza de que han sido escritos por esta persona o se desea saber si dos autores son la misma persona, estas palabras claves podrían aportar algún indicio de ello. Este esquema resulta útil para datos que provienen de fuentes no normalizadas, ya que si la fuente de datos o motor provee algún dato normalizado y obligatorio, por ejemplo el nombre de la universidad o campo de estudio, estos problemas se resolverían de manera más sencilla.

- Otro esquema podría basarse en técnicas de procesamiento del lenguaje natural para hallar o identificar patrones de colaboración entre distintos equipos o conjunto de investigadores y de esta forma resolver el problema de la desambiguación. Del mismo modo, pudiendo analizar el texto resultante de los procesos de recolección, se podrían utilizar listas con nombres geográficos como país y región, o listas de los nombres de universidades y departamentos, para identificar de donde provienen las distintas agrupaciones de autores y así contar con datos adicionales inicialmente no identificados.

- Trabajar con mas de una fuente de datos puede ser algo muy interesante, no solo para obtener una imagen un tanto más "completa" del estado de los resultados de las investigaciones de un científico, sino también para poder comparar estos distintos orígenes, por ejemplo, analizar la cobertura, la cantidad de registros duplicados, la normalización de los datos, la mejora en la normalización de los nombres de autor entre una base de datos y otra.

- Si bien la herramienta desarrollada funciona para búsquedas por autor por la naturaleza del análisis, no requeriría muchos cambios para adaptar dicho análisis y procesamiento de datos, para poder estudiar las publicaciones de instituciones.

- Sería interesante además de resolver el problema de la desambiguación para dos autores distintos, tener la posibilidad de comparar los resultados de ambos, no solo mediante los indicadores planteados sino también a través de la visualización.

Como se expresó quedan abiertas varias líneas y cursos de acción no solo para mejorar lo planteado sino para ampliarlo según las necesidades lo requieran.

6.3. Análisis FODA

A continuación, como punto final, se presenta en la Figura 6.1 el análisis FODA (Fortalezas Oportunidades Debilidades y Amenezas) del modelo diseñado e implementado en esta tesis.

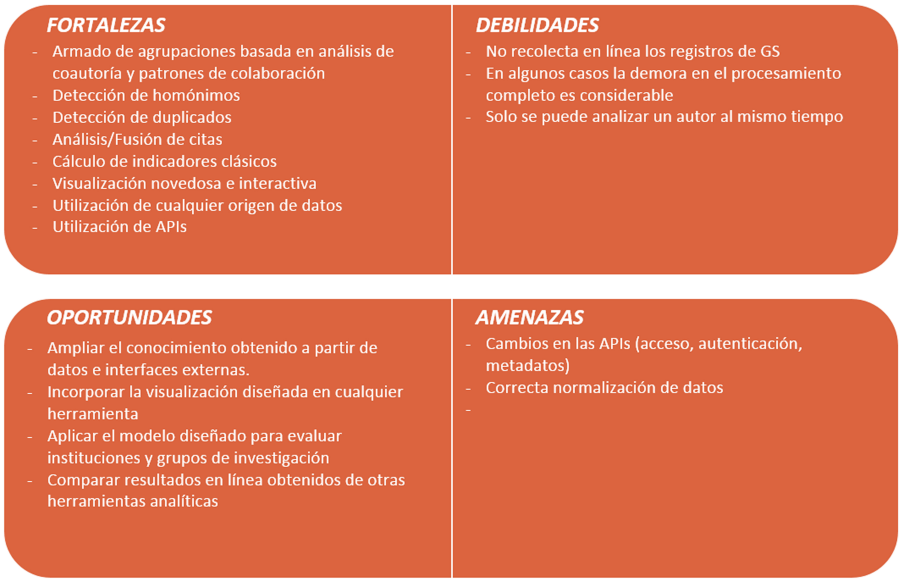


Figura 6.1: Matriz FODA del modelo presentado

Apéndice A

Abreviaturas

AE	Academic Evaluator
aIC	author-centric implicit coauthors
AIS	Article Influence Score
ALM	Article Level Metrics
API	Application Programming Interface
ASNSs	Academic Social Networking services
BASE	Bielefeld Academic Search Engine
CCS	Cascading Style Sheets
cIC	coauthor-centric implicit coauthors
CNKI	China National Knowledge Infrastructure
CWTS	Centre for Science and Technology Studies
DCI	Data Citation Index

DDC	Dewey Decimal Classification
DNS	Domain Name System
DOI	Digital Object Identifier
EIDR	Entertainment Identidier Registry
GB	Giga Byte
GRID	Global Research Identifier Database
GS	Google Scholar
GSACT	Google Scholar Author Citation Tracker
GSC	Google Scholar Citacions
GSM	Google Scholar Metrics
HCI	Human-Computer Interaction
HTML	HyperText Markup Language
IDF	International DOI Foundation
IF	Impact Factor
IMRYD	Introducción, , Metodología, Resultados y Discusión
ISBN	International Standard Book Number
ISI	Institute for Scientific Information
ISNI	International Standard Name Identifier
ISSN	International Standard Serial Number
ISTIC	The Institute of Scientific and Technical Information of China

JaLC	Japan Link Center
JCR	Journal Citation Reports
JIF	Journal Impact Factor
JSON	JavaScript Object Notation
KISTI	Korea Institute of Science and Technology Information
MA	Microsoft Academic
MAG	Microsoft Academic Graph
MAS	Microsoft Academic Search
MC	Mixed Citation
mEDRA	Multilingual European DOI Registration Agency
OA	Open Access
OAI	Open Archives Initiative
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OC	Open Content
OER	Open Educational Resources
OP	Publications Office of the European Union
ORCID	Open Researcher and Contributor ID
PDF	Portable Document Format
PID	Persistent IDentifiers

PLoS	Public Library of Science
PoP	Publish or Perish
RDCP	Relative database citation potential
RIP	Raw Impact per Paper
RSDC	Redes Sociales Digitales Científicas
SC	Split Citation
SJR	Scimago Journal and Country Rank
SNIP	Source Normalized Impact per Paper
SQL	Structured Query Language
SSMS	SQL Server Management Studio
TF-IDF	Term Frequency-Inverse Document Frequency
URL	Uniform Resource Locator
URN	Uniform Resource Name
WoK	Web of Knowledge
WoS	Web of Science

Apéndice B

Stop Words

Listado de palabras vacías empleadas para filtrar los títulos de las publicaciones para realizar consultas a la AK API.

a	p	die	seid	en
about	part	diejenige	seien	encore
above	parted	diejenigen	sein	entre
abstract	particular	dies	seine	envers
across	parting	diese	seinem	environ
after	parts	dieselbe	seinen	es
again	per	dieselben	seiner	ès
against	perhaps	diesem	seines	est
all	place	diesen	seit	et
almost	places	dieser	seitdem	etant
alone	point	dieses	selbst	étaient

along
already
also
although
always
among
an
analysis
analyzed
and
another
any
anybody
anyone
anything
anywhere
are
area
areas
pointed
pointing
points
possible
potentially
present
presented
presenting
presents
problem
problems
produced
proposed
provided
provides
put
puts
q
quite
dir
doch
dort
drei
drin
dritte
dritten
dritter
drittes
du
durch
durchaus
dürfen
dürft
durfte
durften
eben
ebenso
ehrlich
sich
sie
sieben
siebente
siebenten
siebenter
siebentes
sind
so
solang
solche
solchem
solchen
solcher
solches
soll
sollen
sollte
sollten
étais
était
étant
etc
été
etre
être
eu
euh
eux
eux-mêmes
excepté
façon
fais
faisaient
faisant
fait
feront
fi

around	r	ei	sondern	flac
as	rather	eigen	sonst	floc
ask	really	eigene	sowie	font
asked	recent	eigenen	später	gens
asking	related	eigener	statt	ha
asks	report	eigenes	tag	hé
associated	reported	ein	tage	hein
at	required	einander	tagen	hélas
available	result	eine	tat	hem
away	results	einem	teil	hep
b	right	einen	tel	hi
back	right	einer	tritt	ho
backed	room	eines	trotzdem	holà
backing	rooms	einige	tun	hop
backs	s	einigen	über	hormis
based	said	einiger	überhaupt	hors
be	same	einiges	übrigens	hou
became	saw	einmal	uhr	houp
because	say	einmal	um	hue
become	says	eins	und	hui

becomes
been
before
began
behind
being
beings
best
better
between
big
both
but
by
c
came
can
cannot
case
second
seconds
see
seem
seemed
seeming
seems
sees
seven
several
shall
she
should
show
showed
showing
shows
side
sides
elf
en
ende
endlich
entweder
entweder
er
ernst
erst
erste
ersten
erster
erstes
es
etwa
etwas
euch
früher
fünf
uns
unser
unsere
unserer
unter
vergangenen
viel
viele
vielem
vielen
vielleicht
vier
vierte
vierten
vierter
viertes
vom
von
vor
huit
huitième
hum
hurrah
il
ils
importe
je
jusqu
jusque
k
la
là
laquelle
las
le
lequel
les
lès

cases	since	fünfte	wahr	lesquelles
certain	six	fünften	während	lesquels
certainly	small	fünfter	währenddem	leur
clear	smaller	fünftes	währenddessen	leurs
clearly	smallest	für	wann	longtemps
come	so	gab	war	lorsque
compared	some	ganz	wäre	lui
considered	somebody	ganze	waren	lui-même
could	someone	ganzen	wart	ma
d	something	ganzer	warum	maint
demonstrate	somewhere	ganzes	was	mais
demonstrated	state	gar	wegen	malgré
described	states	gedurft	weil	me
did	still	gegen	weit	même
differ	study	gegenüber	weiter	mêmes
different	such	gehabt	weitere	merci
differently	suggest	gehen	weiteren	mes
discussed	sure	geht	weiteres	mien
do	t	gekannt	welche	mienne
does	take	gekonnt	welchem	miennes

done
down
down
downed
downing
downs
due
during
e
each
early
eight
either
end
ended
ending
ends
enough
establish
taken
ten
than
that
the
their
them
then
there
therefore
these
they
thing
things
think
thinks
this
those
though
gemacht
gemocht
gemusst
genug
gerade
gern
gesagt
gesagt
geschweige
gewesen
gewollt
geworden
gibt
ging
gleich
gross
groß
grosse
große
welchen
welcher
welches
wem
wen
wenig
wenig
wenige
weniger
weniges
wenigstens
wenn
wenn
wer
werde
werden
werdet
wessen
wie
miens
mille
mince
moi
moi-même
moins
mon
moyennant
na
ne
néanmoins
neuf
neuvième
ni
nombreuses
nombreux
non
nos
notre

established
establishes
evaluated
even
evenly
ever
every
everybody
everyone
everything
everywhere
f
face
faces
fact
facts
far
felt
few
find
thought
thoughts
three
through
thus
to
today
together
too
took
toward
turn
turned
turning
turns
two
u
under
until
up
grossen
großen
grosser
großer
grosses
großes
gut
gute
guter
gutes
habe
haben
habt
hast
hat
hatte
hätte
hatten
hätten
heisst
wie
wieder
will
willst
wir
wird
wirklich
wirst
wo
wohl
wollen
wollt
wollte
wollten
worden
wurde
würde
wurden
würden
zehn
nôtre
nôtres
nous
nous-mêmes
nul
o
ô
oh
ohé
olé
ollé
on
ont
onze
onzième
ore
ou
où
ouf
ouias

findings
finds
first
five
for
four
from
full
fully
further
furthered
furthering
furthers
g
gave
general
generally
get
gets
upon
us
use
used
uses
using
v
various
very
w
want
wanted
wanting
wants
was
way
ways
we
well
her
heute
hier
hin
hinter
hoch
ich
ihm
ihn
ihnen
ihr
ihre
ihrem
ihren
ihrer
ihres
im
immer
in
zehnte
zehnten
zehnter
zehntes
zeit
zu
zuerst
zugleich
zum
zum
zunächst
zur
zurück
zusammen
zwanzig
zwar
zwar
zwei
zweite
oust
ouste
outre
paf
pan
par
parmi
partant
particulier
particulière
particulièrement
pas
passé
pendant
personne
peu
peut
peuvent
peux

give
given
gives
go
going
good
goods
got
great
greater
greatest
group
grouped
grouping
groups
h
had
has
have
having
wells
went
were
what
when
where
whether
which
whichever
while
who
whole
whose
why
will
with
within
without
work
worked
indem
infolgedessen
ins
irgend
ist
ja
jahr
jahre
jahren
je
jede
jedem
jeden
jeder
jedermann
jedermanns
jedoch
jemand
jemandem
jemanden
zweiten
zweiter
zweites
zwischen
zwölf
à
â
abord
afin
ah
ai
aie
ainsi
allaient
allo
allô
allons
après
assez
attendu
pff
pfft
pfut
pif
plein
plouf
plus
plusieurs
plutôt
pouah
pour
pourquoi
premier
première
premièrement
près
proche
psitt
puisque
qu

he	working	jene	au	quand
her	works	jenem	aucun	quant
here	would	jenen	aucune	quanta
herself	x	jener	aujourd	quant-à-soi
high	y	jenes	aujourd'hui	quarante
high	year	jetzt	auquel	quatorze
high	years	kam	aura	quatre
higher	yet	kann	auront	quatre-vingt
highest	you	kannst	aussi	quatrième
him	young	kaum	autre	quatrièmement
himself	younger	kein	autres	que
his	youngest	keine	aux	quel
how	your	keinem	auxquelles	quelconque
however	yours	keinen	auxquels	quelle
i	z	keiner	avaient	quelles
if	ab	kleine	avais	quelque
ii	aber	kleinen	avait	quelques
iii	aber	kleiner	avant	quelqúun
important	ach	kleines	avec	quels

improve
improved
in
including
increased
interest
interested
interesting
interests
into
is
it
its
itself
j
just
k
keep
keeps
kind

acht
achte
achten
achter
achtes
ag
alle
allein
allem
allen
aller
allerdings
alles
allgemeinen
als
als
also
am
an
andere

kommen
kommt
können
könnt
konnte
könnte
konnten
kurz
lang
lange
lange
leicht
leide
lieber
los
machen
macht
machte
mag
magst

avoir
ayant
bah
beaucoup
bien
bigre
boum
bravo
brrr
ça
car
ce
ceci
cela
celle
celle-ci
celle-là
celles
celles-ci
celles-là

qui
quiconque
quinze
quoi
quoique
revoici
revoilà
rien
sa
sacrebleu
sans
sapristi
sauf
se
seize
selon
sept
septième
sera
seront

knew
know
known
knows
l
large
largely
last
later
latest
least
less
let
lets
like
likely
long
longer
longest
anderen
andern
anders
au
auch
auch
auf
aus
ausser
außer
ausserdem
außerdem
bald
bei
beide
beiden
beim
beispiel
bekannt
mahn
man
manche
manchem
manchen
mancher
manches
mann
mehr
mein
meine
meinem
meinen
meiner
meines
mensch
menschen
mich
mir
celui
celui-ci
celui-là
cent
cependant
certain
certaine
certaines
certains
certes
ces
cet
cette
ceux
ceux-ci
ceux-là
chacun
chaque
cher
ses
si
sien
sienne
siennes
siens
sinon
six
sixième
soi
soi-même
soit
soixante
son
sont
sous
stop
suis
suivant

m	bereits	mit	chère	sur
made	besonders	mittel	chères	surtout
make	besser	mochte	chers	ta
making	besten	möchte	chez	tac
man	bin	mochten	chiche	tant
many	bis	mögen	chut	te
may	bisher	möglich	ci	té
me	bist	mögt	cinq	tel
member	da	morgen	cinquantaine	telle
members	dabei	muss	cinquante	tellement
men	dadurch	muß	cinquantième	telles
method	dafür	müssen	cinquième	tels
might	dagegen	musst	clac	tenant
more	daher	müsst	clic	tes
moreover	dahin	musste	combien	tic
most	dahinter	mussten	comme	tien
mostly	damals	na	comment	tienne
mr	damit	nach	compris	tiennes
mrs	danach	nachdem	concernant	tiens
much	daneben	nahm	contre	toc

must	dank	natürlich	couic	toi
my	dann	neben	crac	toi-même
myself	daran	nein	da	ton
n	darauf	neue	dans	touchant
near	daraus	neuen	de	toujours
necessary	darf	neun	debout	tous
need	darfst	neunte	dedans	tout
needed	darin	neunten	dehors	toute
needing	darüber	neunter	delà	toutes
needs	darum	neuntes	depuis	treize
never	darunter	nicht	derrière	trente
new	das	nicht	des	très
new	das	nichts	dès	trois
newer	dasein	nie	désormais	troisième
newest	daselbst	niemand	desquelles	troisièmement
next	dass	niemandem	desquels	trop
nine	daß	niemanden	dessous	tsoin
no	dasselbe	noch	dessus	tsouin
nobody	davon	nun	deux	tu

non
noone
not
nothing
now
nowhere
number
numbers
o
obtained
of
off
often
old
older
oldest
on
once
one
only
davor
dazu
dazwischen
dein
deine
deinem
deiner
dem
dementsprechend
demgegenüber
demgemäss
demgemäß
demselben
demzufolge
den
denen
denn
denn
denselben
der
nun
nur
ob
oben
oder
offen
oft
oft
ohne
Ordnung
recht
rechte
rechten
rechter
rechtes
richtig
rund
sa
sache
sagt
deuxième
deuxièmement
devant
devers
devra
différent
différente
différentes
différents
dire
divers
diverse
diverses
dix
dix-huit
dixième
dix-neuf
dix-sept
doit
doivent
un
une
unes
uns
va
vais
vas
vé
vers
via
vif
vifs
vingt
vivat
vive
vives
vlan
voici
voilà
vont

open	deren	sagte	donc	vos
opened	derjenige	sah	dont	votre
opening	derjenigen	satt	douze	vôtre
opens	dermassen	schlecht	douzième	vôtres
or	dermaßen	Schluss	dring	vous
order	derselbe	schon	du	vous-mêmes
ordered	derselben	sechs	duquel	vu
ordering	des	sechste	durant	zut
orders	deshalb	sechsten	effet	del
other	desselben	sechster	eh	el
others	dessen	sechstes	elle	las
our	deswegen	sehr	elle-même	los
out	d.h	sei	elles	una
over	dich	sei	elles-mêmes	

Apéndice C

Visualizaciones de datos para otros investigadores

En este apéndice se ofrecen algunos ejemplos de visualizaciones para otros autores fuera de los tenido en cuenta en el desarrollo del trabajo, dichas visualizaciones se obtuvieron utilizando íntegramente la herramienta *Academic Evaluator* y recolectando las publicaciones desde *Google Scholar*.

En las Figuras C.1 y C.2 se observan las visualizaciones tanto antes como luego del análisis para el autor Isidro F. Aguillo. Del mismo modo en las Figuras C.3 y C.4 se observan las visualizaciones tanto antes como luego del análisis para el autor Emilio Delgado López Cózar. Así mismo, en las Figuras C.5 y C.6 se observan las visualizaciones tanto antes como luego del análisis para la autora Anne Wil Harzing. También, en las Figuras C.7 y C.8 se observan las visualizaciones tanto antes como luego del análisis para el autor Jason Priem. Y por último, en las Figuras C.9 y C.10 se observan las visualizaciones tanto antes como luego del análisis para el autor Daniel Torres Salinas.

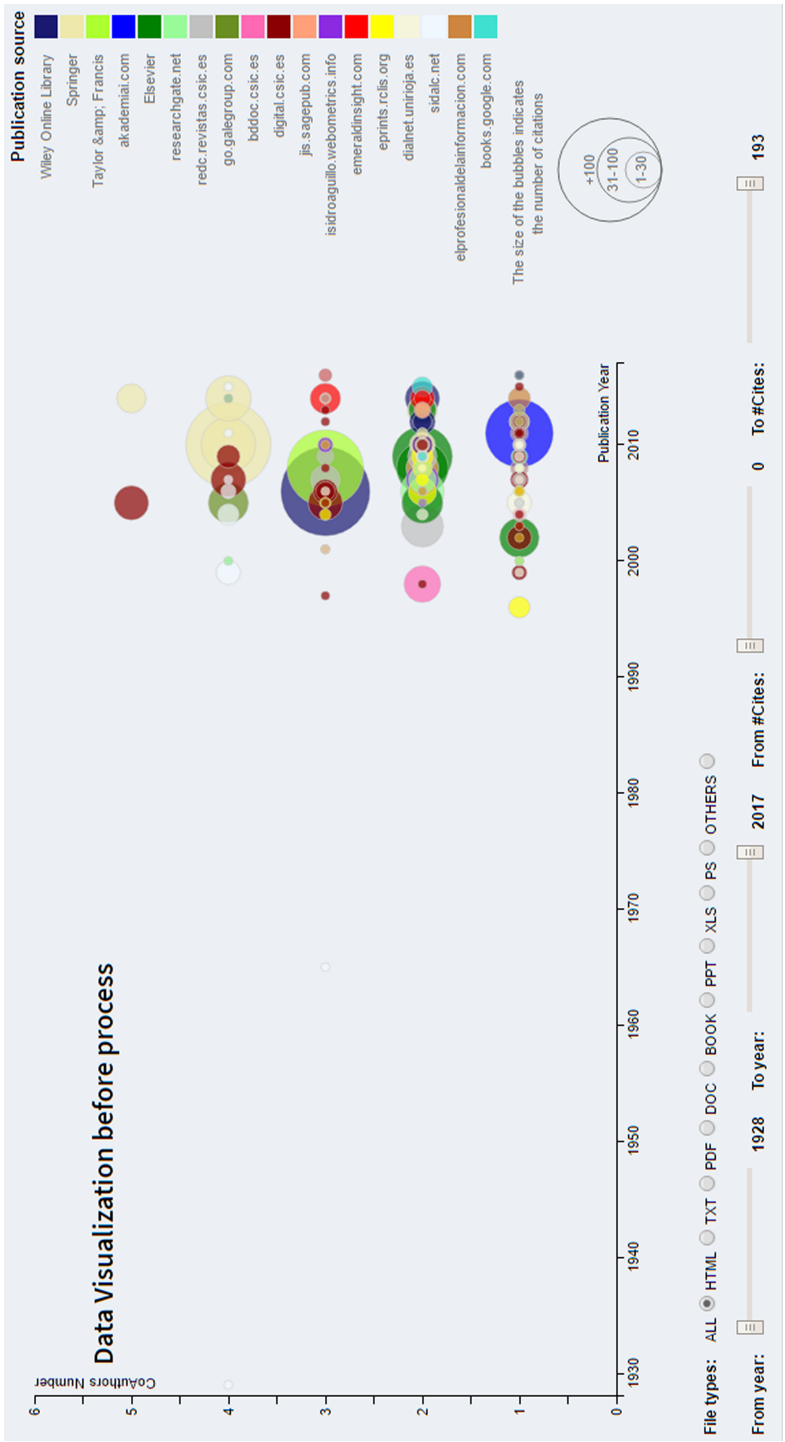


Figura C.1: Visualización de información antes del análisis para el autor Isidro F. Aguillo, *source*=GS

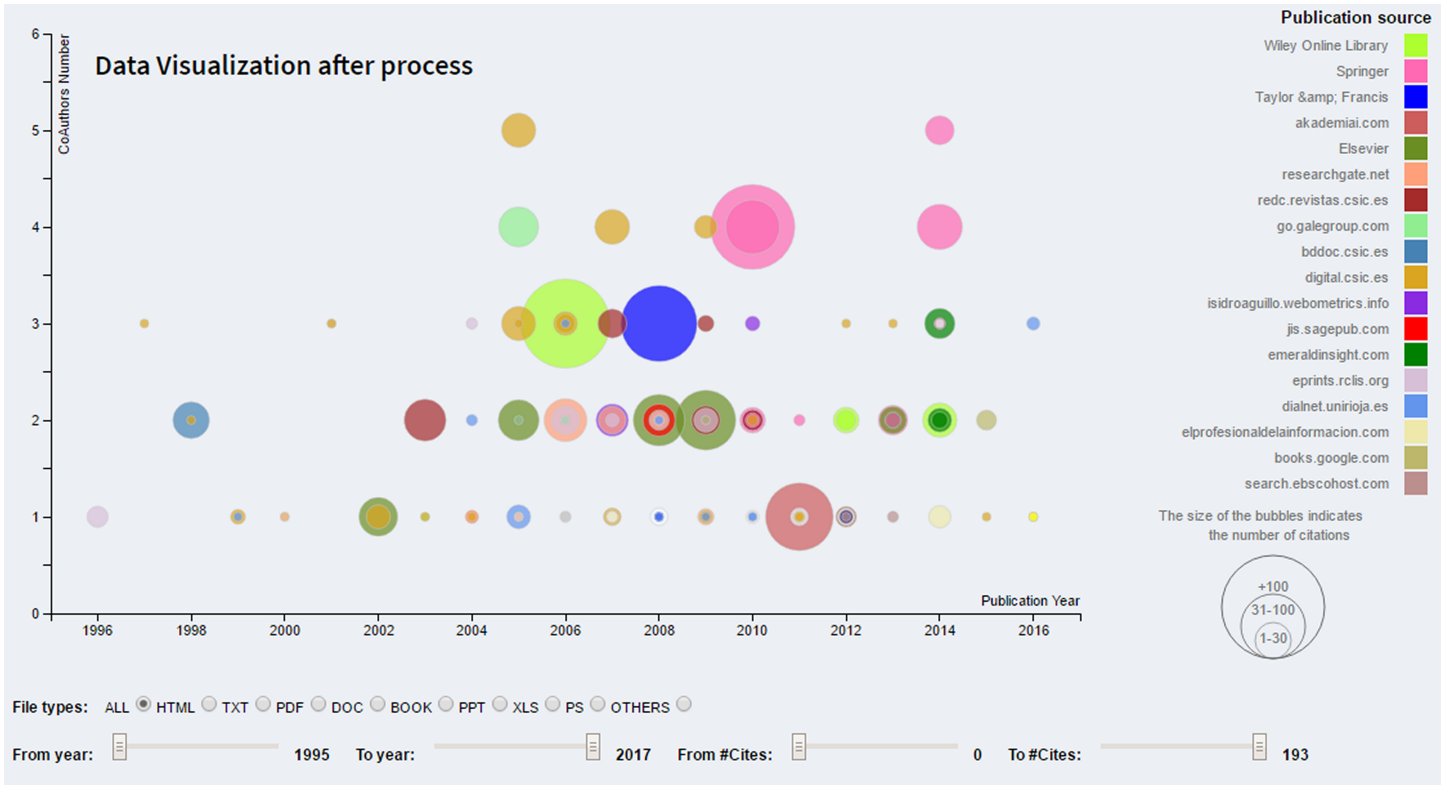


Figura C.2: Visualización de información luego del análisis para el autor Isidro F. Aguillo, *source*=GS

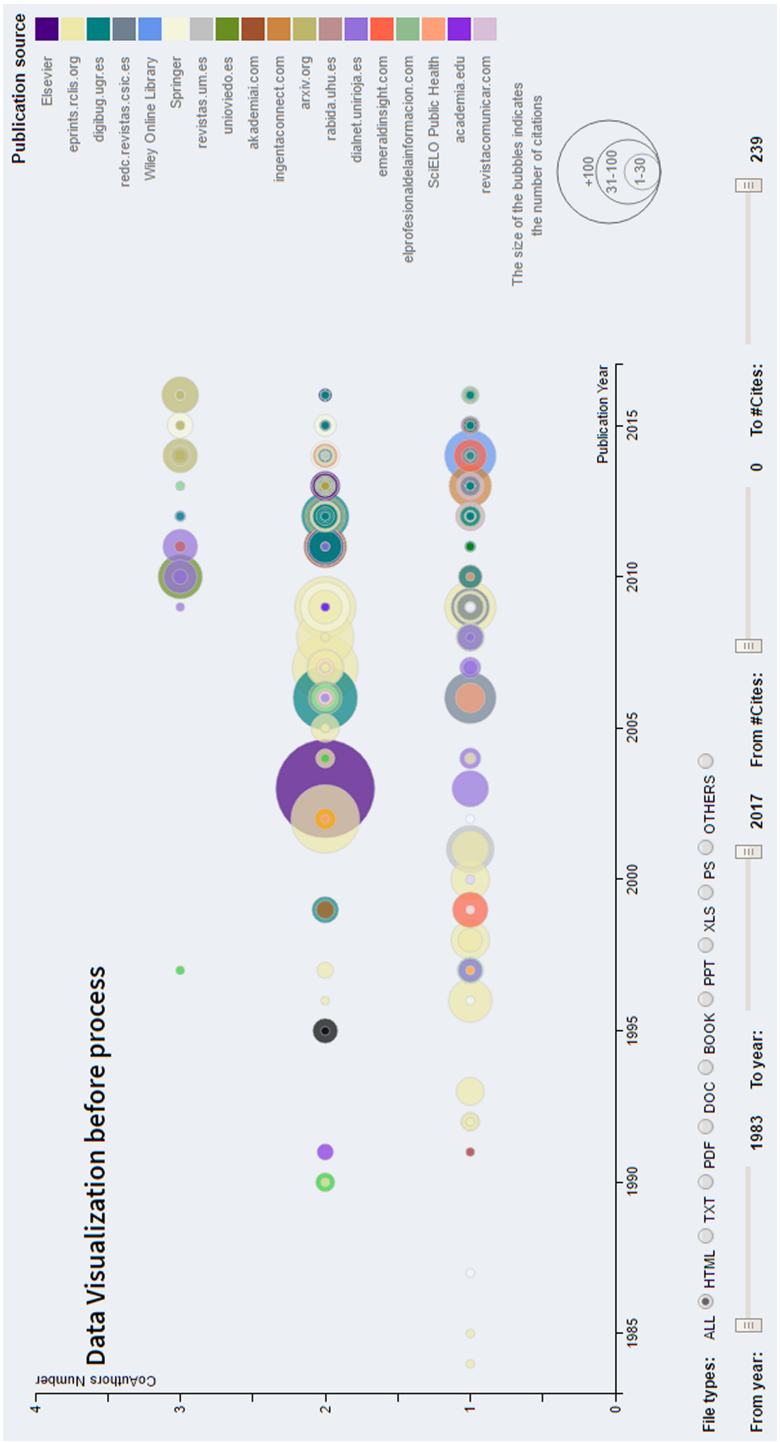


Figura C.3: Visualización de información antes del análisis para el autor Emilio Delgado López Cózar, *source*=GS

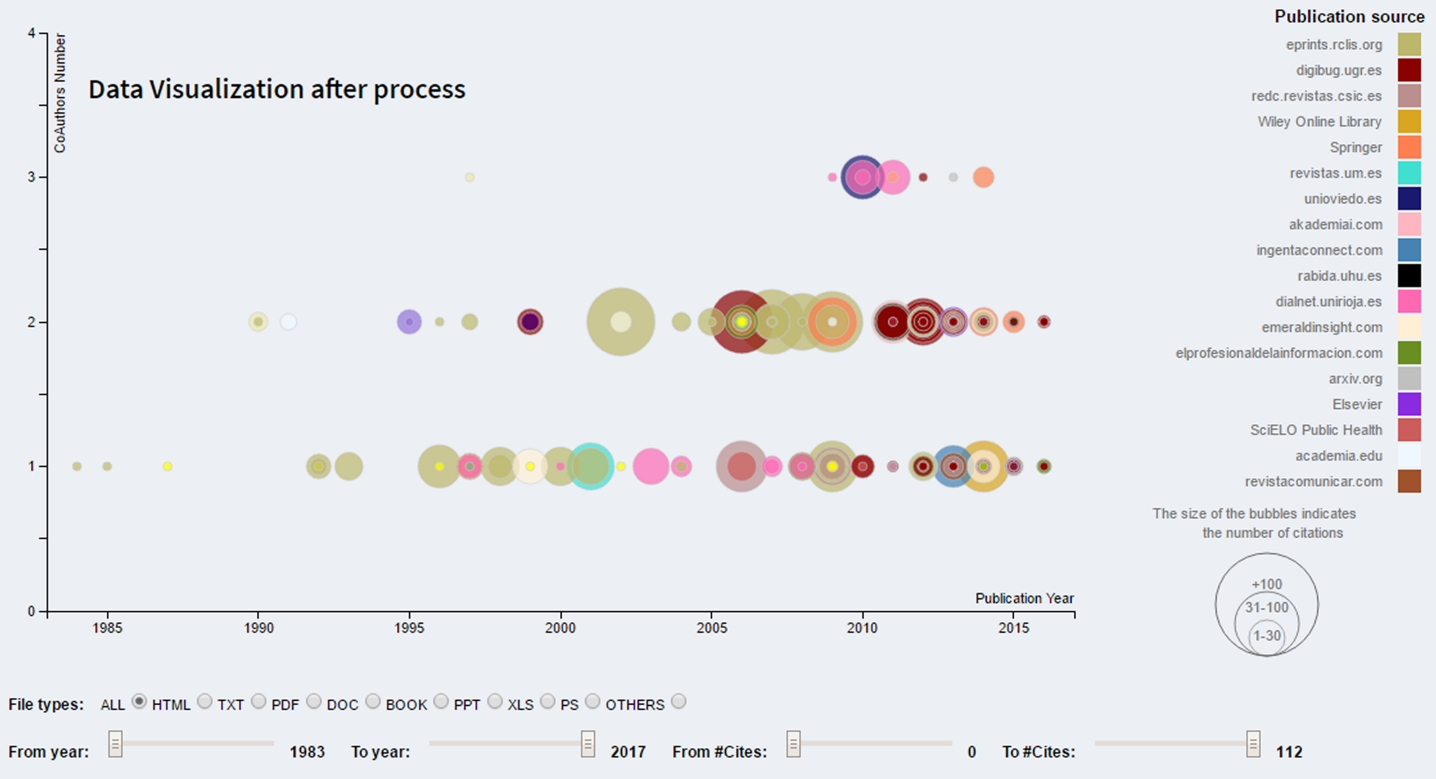


Figura C.4: Visualización de información luego del análisis para el autor Emilio Delgado López Cózar, *source*=GS

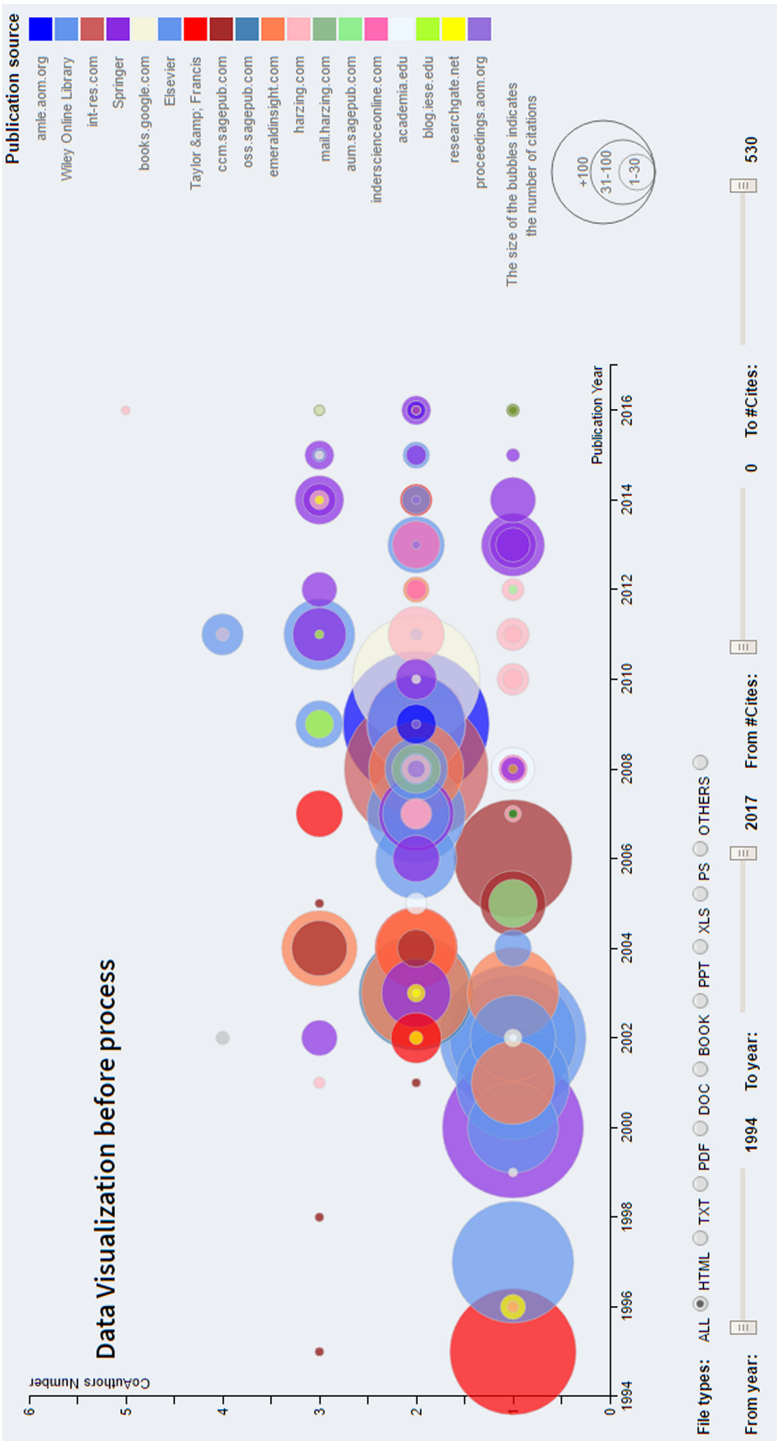


Figura C.5: Visualización de información antes del análisis para la autora Anne Wil Harzing, *source*=GS

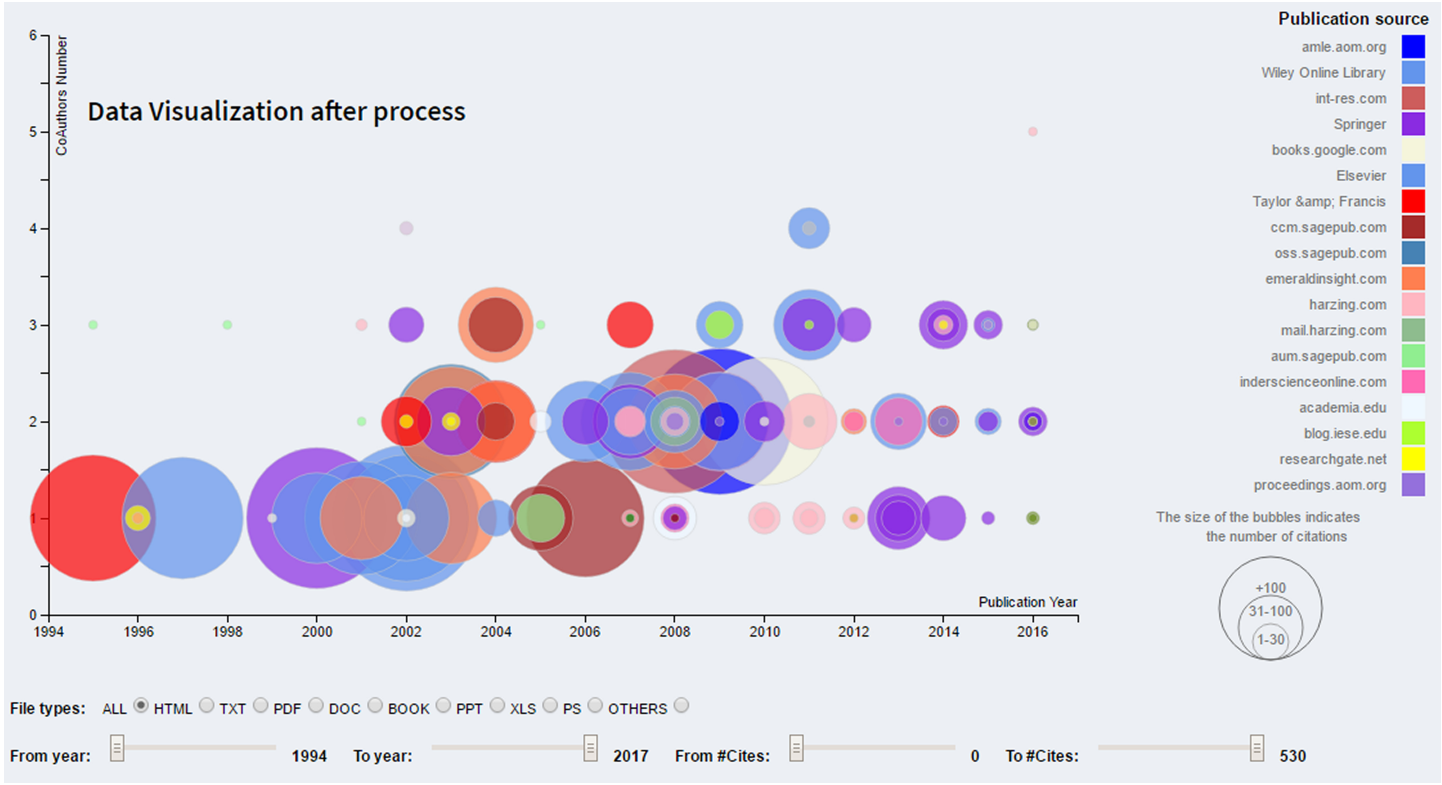


Figura C.6: Visualización de información luego del análisis para la autora Anne Wil Harzing, *source*=GS

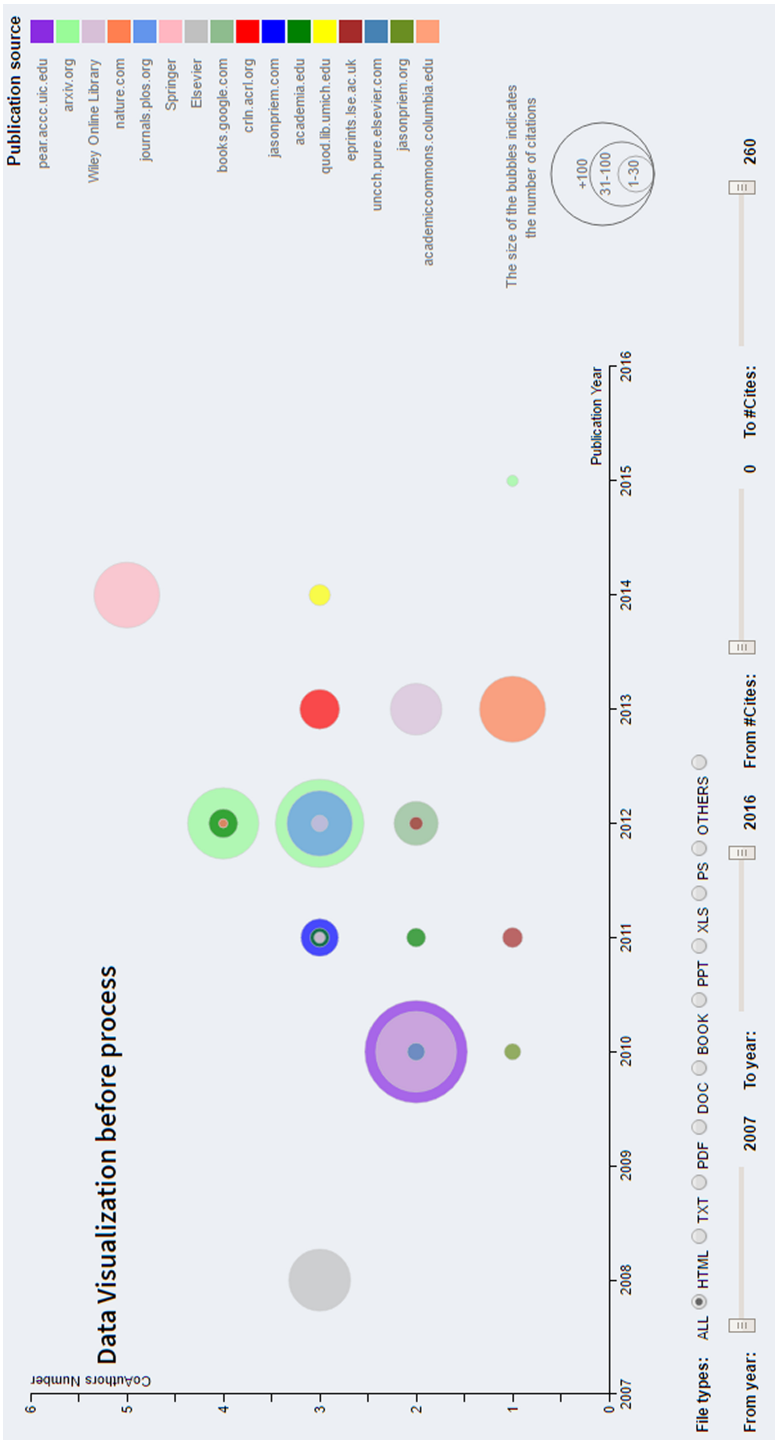


Figura C.7: Visualización de información antes del análisis para el autor Jason Priem, *source*=GS

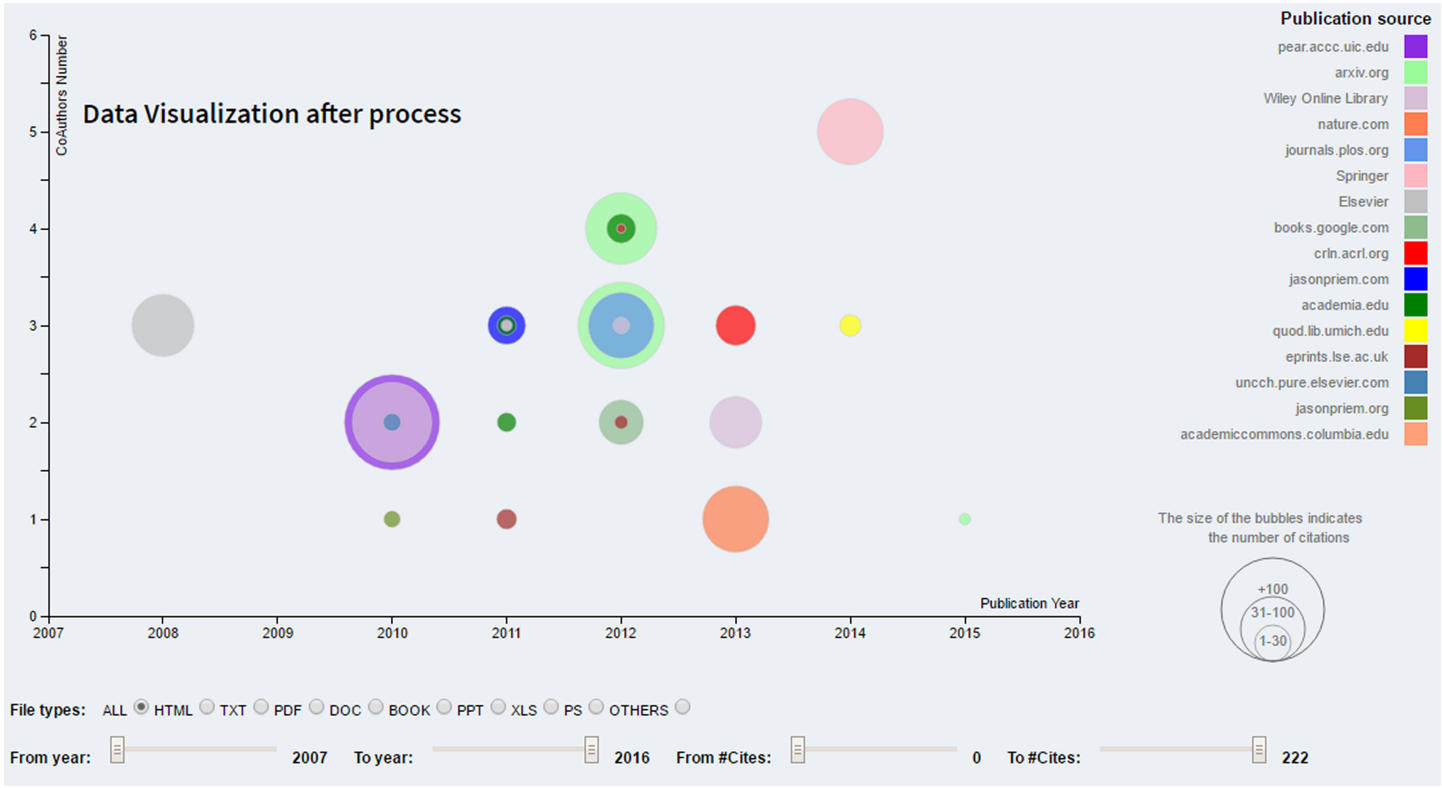


Figura C.8: Visualización de información luego del análisis para el autor Jason Priem, *source*=GS

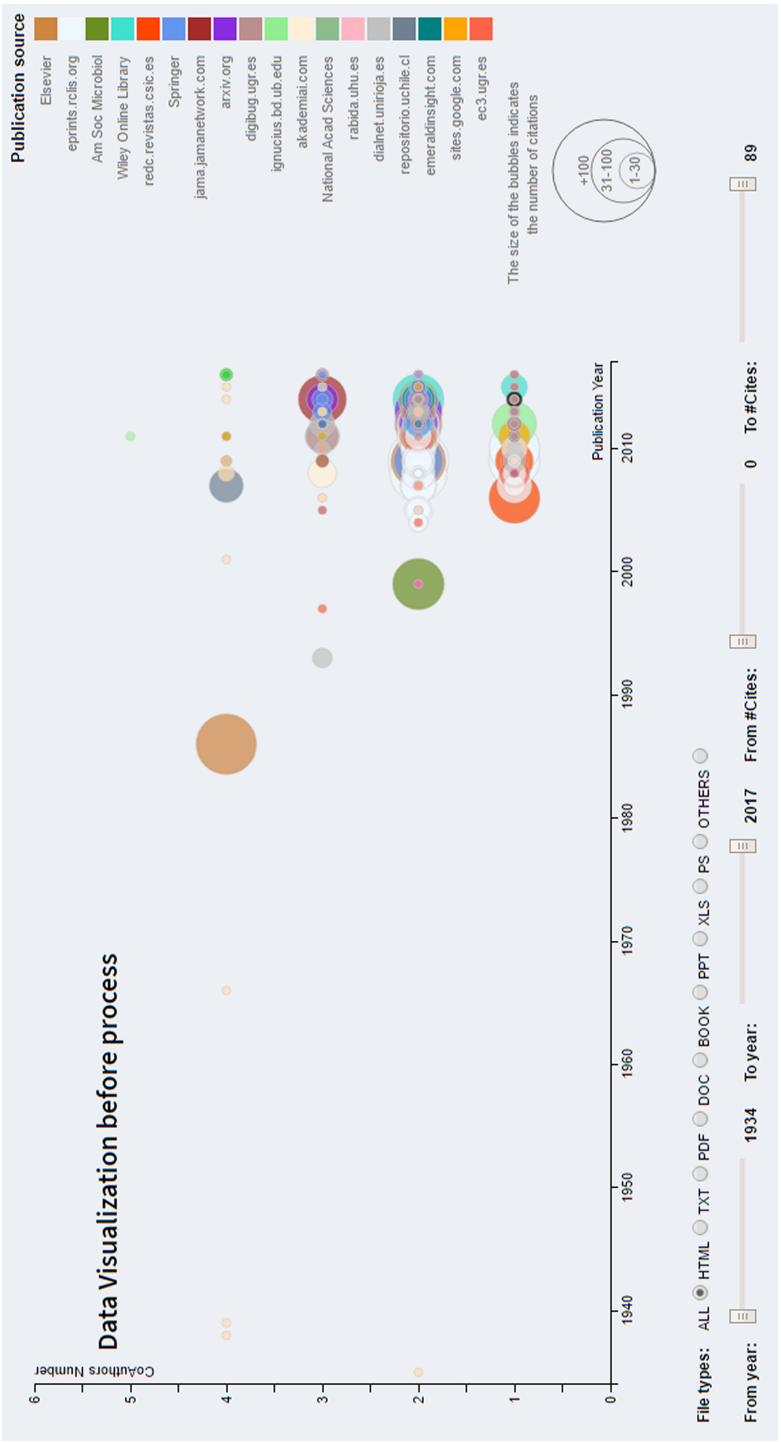


Figura C.9: Visualización de información antes del análisis para el autor Daniel Torres Salinas, *source*=GS

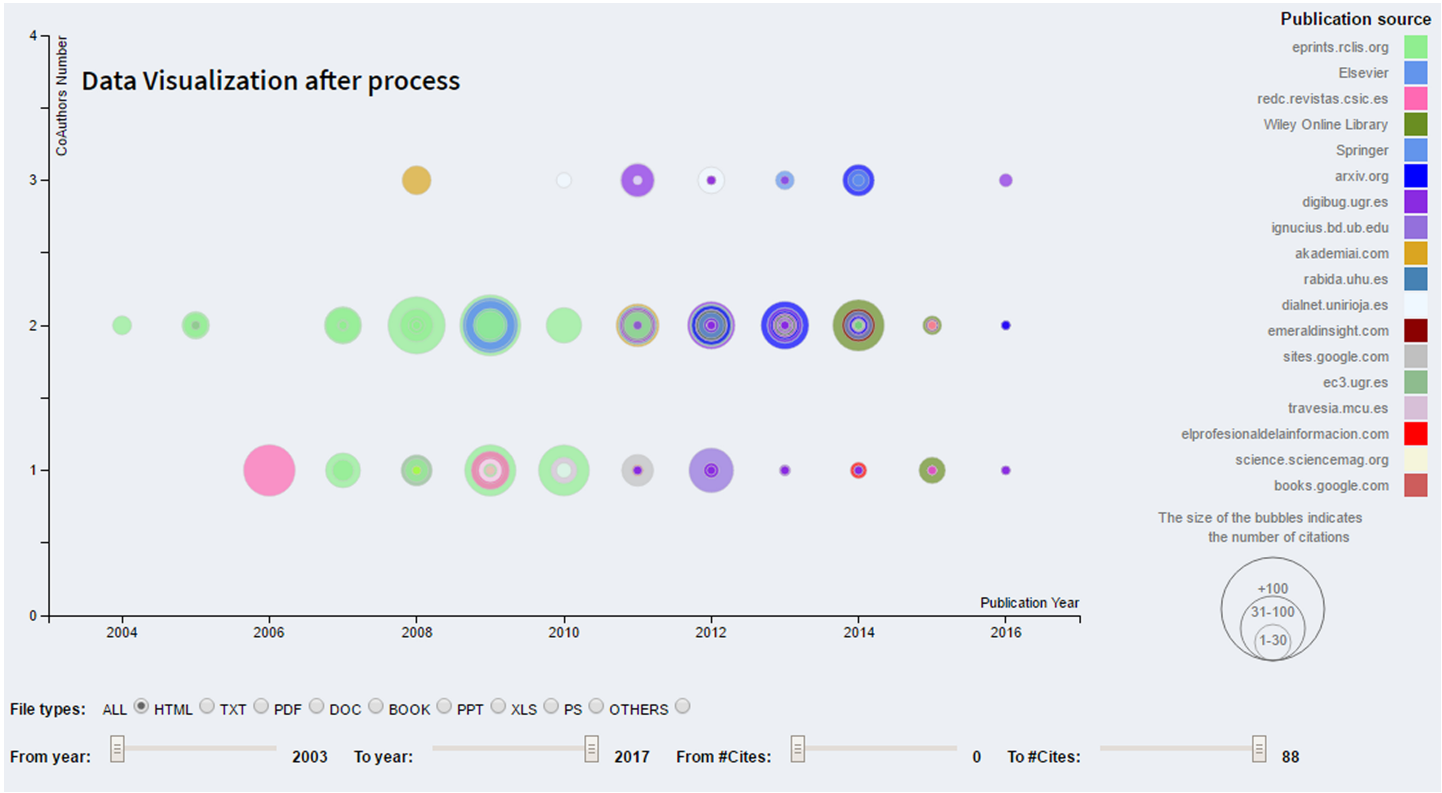


Figura C.10: Visualización de información luego del análisis para el autor Daniel Torres Salinas, *source*=GS

Bibliografía

Abdulhayoglu, M. A. and Thijs, B. (2017). Use of researchgate and google cse for author name disambiguation. *Scientometrics*, pages 1–21.

Abdulkhudhur, H. N., Habeeb, I. Q., Yusof, Y., and Yusof, S. A. M. (2016). Implementation of improved levenshtein algorithm for spelling correction word candidate list generation. *Journal of Theoretical and Applied Information Technology*, 88(3):449–455.

Abernethy, J., Chapelle, O., and Castillo, C. (2010). Graph regularization methods for Web spam detection. *Mach. Learn.*, 81(2):207–225.

Academia.edu (2016). About academia.edu. Disponible en: `https://www.academia.edu/about` Visitado el: 09/10/2016.

Adler, R., Ewing, J., and Taylor, P. (2009). Citation statistics. *Statistical Science*, 24(1):1–14.

Aghaei Chadegani, A., Salehi, H., Yunus, M. M., Farhadi, H., Fooladi, M., Farhadi, M., and Ale Ebrahim, N. (2013). A Comparison between Two Main Academic Literature Collections: Web of Science and Scopus Databases. *Asian Social Science*, 9(5):18–26.

Aguillo, I. F. (2012). Is google scholar useful for bibliometrics? a webometric analysis. *Scientometrics*, 91:343–351.

Aigner, W., Miksch, S., Müller, W., Schumann, H., and Tominski, C. (2007). Visualizing time-oriented data-a systematic view. *Comput. Graph.*, 31(3):401–409.

Aizawa, A. (2003). An information-theoretic perspective of tf-idf measures. *Information Processing & Management*, 39(1):45–65.

Akers, K. G., Sarkozy, A., Wu, W., and Slyman, A. (2016). Orcid author identifiers: A primer for librarians. *Medical Reference Services Quarterly*, 35(2):135–144.

Albo, Y., Lanir, J., Bak, P., and Rafaeli, S. (2016). Off the radar: Comparative evaluation of radial visualization solutions for composite indicators. *IEEE Trans. Vis. Comput. Graph.*, 22(1):569–578.

Alemasoom, H., Samavati, F. F., Brosz, J., and Layzell, D. (2014). Interactive visualization of energy system. In *Cyberworlds (CW), 2014 International Conference on*, pages 229–236. IEEE.

Allison, P. D. and Steward, J. A. (1974). Productivity differences among scientists: evidence for accumulative advantage. *American Sociological Review*, 39(4):596–606.

Alonso, S., Cabrerizo, F., Herrera-Viedma, E., and Herrera, F. (2010). hg-index: a new index to characterize the scientific output of researchers based on the h- and g-indices. *Scientometrics*, 82:391–400.

Altmetric.com (2016). The donut and altmetric attention score. Disponible en: https://www.altmetric.com/about-our-data/the-donut-and-score/ Visitado el: 01/11/2016.

AltmetricSupport (2016). What outputs and sources does altmetric track? Disponible en: https://help.altmetric.com/support/solutions/articles/6000060968-what-outputs-and-sources-does-altmetric-track- Visitado el: 09/10/2016.

Andalia, R. C., Rodríguez, M. N., and Rodríguez, K. M. P. (2015). Orcid: en busca de un identificador único permanente y universal para científicos y académicos. *Revista Cubana de Información en Ciencias de la Salud*, 26(1):71–77.

Andrews, K. (2011a). Human-computer interaction. Lecture notes, Graz University of Technology.

Andrews, K. (2011b). Information visualisation. Course notes, Graz University of Technology.

Arévalo, J. A., Cordón-Garcia, J. A., and Barba, B. M. (2016). Altmetrics: medición de la influencia de los medios en el impacto social de

la investigación. *Cuadernos de Documentación Multimedia*, 27(1):75–101.

Arévalo, J. A. and Vázquez, M. V. (2016). Altmetrics y alfabetización científica. *bibliotecas anales de investigación*, 12(1):14–29.

AV, K., B, A., I, S., and JW, B. (2009). Comparisons of citations in web of science, scopus, and google scholar for articles published in general medical journals. *JAMA*, 302(10):1092–1096.

Bach, B., Dragicevic, P., Archambault, D., Hurter, C., and Carpendale, S. (2014). A review of temporal data visualizations based on space-time cube operations. *Eurographics Conference on Visualization*.

Bach, B., Dragicevic, P., Archambault, D., Hurter, C., and Carpendale, S. (2016). A descriptive framework for temporal data visualizations based on generalized space-time cubes. *Computer Graphics Forum*, pages n/a–n/a.

Baneyx, A. (2008). "Publish or Perish" as citation metrics used to analyze scientific output in the humanities: International case studies in economics, geography, social sciences, philosophy, and history. *Archivum Immunologiae et Therapiae Experimentalis*, 56(6):363–371.

Bar-Ilan, J. (2008). Which h-index? A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, 74(2):257–271.

BASE (2016a). About base: Statistics. Disponible en: `https://www.base-search.net/about/en/index.php` Visitado el: 21/11/2016.

BASE (2016b). About base: Statistics. Disponible en: `https://www.base-search.net/about/en/about_statistics.php?menu=2` Visitado el: 21/11/2016.

Batista, P., Campiteli, M., and Kinouchi, O. (2006). Is it possible to compare researchers with different scientific interests? *Scientometrics*, 68(1):179–189.

Bayaz, R. (2015). Valuable insight into the broader impacts and dissemination of nearly 180,000 books and more than three million chapters published by springer. Disponible en: `http://tinyurl.com/jz4ftpd` Visitado el: 06/11/2016.

Beel, J. and Gipp, B. (2010). Academic Search Engine Spam and Google Scholar's Resilience Against it. *JOURNAL OF ELECTRONIC PUBLISHING*, 13(3).

Bergstrom, C. (2007). Eigenfactor: Measuring the value and prestige of scholarly journals. *College & Research Libraries News*, 68(5):314–316.

Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P., and Fienberg, S. (2003). Adaptive Name Matching in Information Integration. *IEEE Intelligent Systems*, 18(5):16–23.

Bilenko, M. and Mooney, R. J. (2003). Adaptive Duplicate Detection Using Learnable String Similarity Measures. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003)*, pages 39–48.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag, New York.

Bitton, D. and DeWitt, D. J. (1983). Duplicate record elimination in large data files. *ACM Trans. Database Syst.*, 8:255–265.

Bornmann, L., Mutz, R., and Daniel, H.-D. (2008). Are there better indices for evaluation purposes than the h index? a comparison of nine different variants of the h index using data from biomedicine. *J. Am. Soc. Inf. Sci. Technol.*, 59(5):830–837.

Bostock, M. (2017). Scatterplot. Disponible en: `https://bl.ocks.org/mbostock/3887118`, Visitado el: 02/01/2017.

Braam, R., Moed, H., and Van Raan, A. (1988). Mapping of science : Critical elaboration and new approaches, a case study in agricultural biochemistry. *Egghe, L. & Rousseau, R. (Ed.) Informetrics 87/88, Belgium : Diepenbeek*, pages 15–28.

Branting, L. K. (2002). Name-matching algorithms for legal case-management systems. *Journal of Information, Law and Technology (JILT),*, 1.

Brunnermeier, M. K. (2009). Deciphering the liquidity and credit crunch 2007–2008. *The Journal of economic perspectives*, 23(1):77–100.

Cabezas-Clavijo, A. and Delgado-López-Cózar, E. (2012). ¿es posible usar google scholar para evaluar las revistas científicas nacionales en los ámbitos de ciencias sociales y jurídicas? el caso de las revistas españolas. *EC3 Working Papers*, 3.

Cabezas-Clavijo, A. and Torres-Salinas, D. (2010). Indicadores de uso y participación en las revistas científicas 2.0: el caso de PLoS One. *El profesional de la información*, 19(4):431–434.

Cabezas-Clavijo, A. and Torres-Salinas, D. (2012). Google scholar citations y la emergencia de nuevos actores en la evaluación de la investigación. *Anuario ThinkEPI*, 6.

Canelo, J. A. M., Alonso Sardón, M., Méndez Pardo, M., López León, I., and Sáenz González, M. d. C. (2002). Mortalidad prematura por enfermedades infecciosas en españa, 1908-1995. *Revista Panamericana de Salud Pública*, 12(4).

Capaccioni, A. and Spina, G. (2012). Italian SSH journals in Journal Citation Reports (JCR) and in SCImago Journal Rank (SJR): data and first analysis. *Italian Journal of Library & Information Science*, 3(1):4787–4807.

Card, S., Mackinlay, J., and Shneiderman, B. (1999). *Readings in Information Visualization: Using Vision to Think*. Morgan-Kaufmann, San Francisco.

Chacón, A., Marco-Sola, S., Espinosa, A., Ribeca, P., and Moure, J. C. (2014). Thread-cooperative, bit-parallel computation of levenshtein distance on gpu. In *Proceedings of the 28th ACM International Conference on Supercomputing*, ICS '14, pages 103–112, New York, NY, USA. ACM.

Chaudhuri, S., Ganti, V., and Kaushik, R. (2006). A Primitive Operator for Similarity Joins in Data Cleaning. In *ICDE '06: Proceedings of the 22nd International Conference on Data Engineering*, page 5, Washington, DC, USA. IEEE Computer Society.

Chen, X. (2010). Google Scholar's Dramatic Coverage Improvement Five Years after Debut. *Serials Review*, 36(4):221–226.

Chum, O., Philbin, J., and Zisserman, A. (2008). Near Duplicate Image Detection: min-Hash and tf-idf Weighting. *Proceedings of the 19th British Machine Vision Conference*, 3:493–502.

CiteUlike (2016). Citeulike is a free service for managing and discovering scholarly references. Disponible en: http://www.citeulike.org/ Visitado el: 09/10/2016.

Clark, R. (2012). What relationship between ISNI and ORCID. Disponible en: https://orcid.org/content/what-relationship-between-isni-and-orcid, Visitado el: 16/02/2017.

Clermont, M. and Dyckhoff, H. (2012). Coverage of business administration literature in google scholar: Analysis and comparison with econbiz, scopus and web of science. *Bibliometrie - Praxis und Forschung*, 1.

Cohen, W., Kautz, H., and McAllester, D. (2000). Hardening soft information sources. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '00, pages 255–259, New York, NY, USA. ACM.

Cohen, W., Ravikumar, P., and Fienberg, S. (2003). A comparison of string distance metrics for name-matching tasks. Workshop on Information Integration on the Web.

Cohen, W. W., Ravikumar, P., Fienberg, S., and Rivard, K. (2013a). Class JaroWinklerTFIDF. Disponible en: http://secondstring.sourceforge.net/javadoc/com/wcohen/ss/JaroWinklerTFIDF.html, Visitado el: 30/04/2017.

Cohen, W. W., Ravikumar, P., Fienberg, S., and Rivard, K. (2013b). Class MongeElkan. Disponible en: http://secondstring.sourceforge.net/javadoc/com/wcohen/secondstring/MongeElkan.html, Visitado el: 30/04/2017.

Colledge, L., de Moya-Anegón, F., Guerrero-Bote, V., López-Illescas, C., El Aisati, M., and Moed, H. (2010). SJR and SNIP: two new journal metrics in Elsevier's Scopus. *Serials: The Journal for the Serials Community*, 23(3):215–221.

Cornell University Library (2017). Measuring your research impact: i10-index. Disponible en: http://guides.library.cornell.edu/c.php?g=32272&p=203393, Visitado el: 07/05/2017.

Costas, R. and Bordons, M. (2007). Algoritmos para solventar la falta de normalización de nombres de autor en los estudios bibliométricos. *Investigación Bibliotecológica*, 21(42).

Dagan, I., Lee, L., and Pereira, F. C. N. (1999). Similarity-Based Models of Word Cooccurrence Probabilities. *Machine Learning*, 34(1):43–69.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176.

Delgado López-Cózar, E., Robinson-Garcia, N., and Torres-Salinas, D. (2014). The google scholar experiment: How to index false papers and manipulate bibliometric indicators. *Journal of the Association for Information Science & Technology*, 65(3):446–454.

Delgado López-Cózar, E. and Caballero, R. R. (2013). The impact of scientific journals of communication: Comparing google scholar metrics, web of science and scopus. *Comunicar*, 21(41):45–52.

Delgado López-Cózar, E., Robinson-García, N., and Torres-Salinas, D. (2012). Manipular google scholar citations y google scholar metrics: simple, sencillo y tentador. *EC3 Working Papers*, 6.

Digital Repository Infrastructure Vision for European Research (2008). Directrices driver 2.0: directrices para proveedores de contenido - exposición de recursos textuales con el protocolo oai-pmh. Disponible en: `http://travesia.mcu.es/portalnb/jspui/handle/10421/1441`, Visitado el: 30/04/2017.

Dixon, L., Duncan, C., Fagan, J. C., Mandernach, M., and Warlick, S. E. (2010). Finding Articles and Journals via Google Scholar, Journal Portals, and Link Resolvers Usability Study Results. *REFERENCE & USER SERVICES QUARTERLY*, 50(2).

Dornberger, R., Hil, D., Wittwer, J., and Burgy, P. (2016). The time diagram control approach for the dynamic representation of time-oriented data. *Systemics, Cybernetics and Informatics*, 14(2):54–60.

Dresden, A. (1922). A report on the scientific work of the chicago section, 1897-1922. *Bulletin of The American Mathematical Society*, 28:303–307.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1999). *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Dutta, B. (2015). Open research data without borders. *Journal of Scientometric Research*, 4(2):122–123.

Egghe, L. (2006a). An improvement of the h-index: the g-index. *ISSI Newsletter*, pages 8–9.

Egghe, L. (2006b). Theory and practise of the g-index. *Scientometrics*, 69(1):131–152.

Elmagarmid, A. K., Ipeirotis, P. G., and Verykios, V. S. (2007). Duplicate Record Detection: A Survey. *Knowledge and Data Engineering, IEEE Transactions on*, 19(1):1–16.

Elsevier (2016a). Elsevier expands metrics perspectives with launch of new altmetrics pilots. Disponible en: `http://tinyurl.com/hhds89v` Visitado el: 02/11/2016.

Elsevier (2016b). What content is included in scopus? Disponible en: `https://www.elsevier.com/solutions/scopus/content` Visitado el: 02/10/2016.

Eltermann, F., Godoy, A., and Zuben, F. J. V. (2016). Effects of social ties in knowledge diffusion: case study on PLOS ONE. *CoRR*, abs/1610.09164.

ePIC (2017). Persistent Identifiers for eResearch Consortium (ePIC). Disponible en: `http://www.pidconsortium.eu/` Visitado el: 16/02/2017.

Factsheet (2015). DOI System and the Handle System. Disponible en: `http://www.doi.org/factsheets/DOIHandle.html`, Visitado el: 16/02/2017.

Falagas, M. E., Kouranos, V. D., Arencibia-Jorge, R., and Karageorgopoulos, D. E. (2008). Comparison of scimago journal rank indicator with journal impact factor. *The FASEB Journal. Life Sciences Forum*, 22(8):2623–2628.

Fellegi, I. P. and Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, 64(328):1183–1210.

Ferreira, A. A., Gonçalves, M. A., and Laender, A. H. (2012). A brief survey of automatic methods for author name disambiguation. *SIGMOD Rec.*, 41(2):15–26.

Fersht, A. (2009). The most influential journals: Impact Factor and Eigenfactor. *Proceedings of the National Academy of Sciences*, 106(17):6883–6884.

Fetterly, D. (2007). Adversarial Information Retrieval: The Manipulation of Web Content. *ACM Computing Reviews*.

Figshare (2016). Figshare about. Disponible en: `https://figshare.com/about` Visitado el: 02/11/2016.

Fox, M. (1983). Publication productivity among scientists: A critical review. *Social Studies of Science*, 13(2):285–305.

García-Gómez, C. (2012). Orcid: un sistema global para la identificación de investigadores. *El profesional de la información*, 21(2):210–212.

García-Peñalvo, F. J., de Figuerola, C. G., and Merlo-Vega, J. A. (2010). Open knowledge: challenges and facts. *Online Information Review*, 34(4):520–539.

Garfield, E. (1972). Citation Analysis as a Tool in Journal Evaluation. *Science*, 178(4060):471–479.

Garfield, E. (1979). *Citation indexing, its theory and application in science, technology, and humanities*. Information sciences series. ISI Press.

Garfield, E. (2006). The History and Meaning of the Journal Impact Factor. *Journal of the American Medical Association*, 295:90–93.

Ginsparg, P., Luce, R., and Van de Sompel, H. (1999). The open archives initiative aimed at the further promotion of author self-archived solutions.

González-Pereira, B., Guerrero Bote, V. P., and de Moya Anegón, F. (2009). The sjr indicator: A new indicator of journals'scientific prestige. *CoRR*, abs/0912.4141.

González-Díaz, C., Iglesias-García, M., and Codina, L. (2015). Presencia de las universidades españolas en las redes sociales digitales científicas: caso de los estudios de comunicación. *El profesional de la información*, 24(5):640–647.

Google (2016). Google scholar metrics. Disponible en: `https://scholar.google.com/intl/en/scholar/metrics.html` Visitado el: 21/11/2016.

Gorbea Portal, S. (2005). *El Modelo Matemático de Lotka: su aplicación a la producción científica latinoamericana en ciencias bibliotecológica y de la información.* México, D.F.: Centro Universitario de Investigaciones Bibliotecológicas, Universidad Nacional Autónoma de México.

Gorraiz, J., Melero-Fuentes, D., Gumpenberger, C., and Valderrama-Zurián, J.-C. (2016). Availability of digital object identifiers (dois) in web of science and scopus. *Journal of Informetrics*, 10(1):98 – 109.

Greenhill, S. J. (2011). Levenshtein distances fail to identify language relationships accurately. *Computational Linguistics*, 37(4):689–698.

GRID (2017a). Disambiguate. resolve unstructured text affiliations to grid ids. Disponible en: `https://www.grid.ac/pages/disambiguate_your_data`, Visitado el: 16/02/2017.

GRID (2017b). Global Research Identifier Database. Disponible en: `https://www.grid.ac/`, Visitado el: 16/02/2017.

GRID (2017). Policies. Disponible en: `https://www.grid.ac/pages/policies`, Visitado el: 16/02/2017.

Groote, S. L. D. and Raszewski, R. (2012). Coverage of google scholar, scopus, and web of science: A case study of the h-index in nursing. *Nursing Outlook*, 60(6):391 – 400. Special Issue: State of the Science: Palliative Care and End of Life.

Gurney, T., Horlings, E., and Van Den Besselaar, P. (2012). Author disambiguation using multi-aspect similarity indicators. *Scientometrics*, 91(2):435–449.

Gyongyi, Z. and Molina, H. G. (2005). Web Spam Taxonomy. In *First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2005)*.

Haak, L. L., Fenner, M., Paglione, L., Pentz, E., and Ratner, H. (2012). Orcid: a system to uniquely identify researchers. *Learned Publishing*, 25(4):259–264.

Haley, M. R. (2014). Ranking top economics and finance journals using microsoft academic search versus google scholar: How does the new publish or perish option compare? *Journal of the Association for Information Science & Technology*, 65(5):1079–1084.

Hamers, L., Hemeryck, Y., Herweyers, G., Janssen, M., Keters, H., Rousseau, R., and Vanhoutte, A. (1989). Similarity measures in scientometric research: The jaccard index versus salton's cosine formula. *Information Processing & Management*, 25(3):315 – 318.

Han, H., Giles, L., Zha, H., Li, C., and Tsioutsiouliklis, K. (2004). Two supervised learning approaches for name disambiguation in author citations. In *JCDL '04: Proceedings of the 4th ACM/IEEE joint conference on Digital libraries*, pages 296–305.

Han, H., Zha, H., and Giles, L. C. (2005). Name disambiguation in author citations using a K-way spectral clustering method. In *JCDL '05: Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pages 334–343, New York, NY, USA. ACM Press.

Harzing, A.-W. (2007). Publish or perish. Disponible en: `http://www.harzing.com/pop.htm`.

Harzing, A.-W. (2010). *The Publish or Perish Book: Your Guide to Effective and Responsible Citation Analysis*. Tarma Software Research.

Harzing, A.-W. (2016a). Microsoft academic (search): A phoenix arisen from the ashes? *Scientometrics*, 108(3):1637–1647.

Harzing, A.-W. (2016b). Publish or perish metrics. Disponible en: `http://www.harzing.com/resources/publish-or-perish` Visitado el: 03/10/2016.

Harzing, A.-W. and Alakangas, S. (2016). Google scholar, scopus and the web of science: a longitudinal and cross-disciplinary comparison. *Scientometrics*, 106(2):787–804.

Harzing, A.-W. and Alakangas, S. (2017). Microsoft Academic: is the phoenix getting wings? *Scientometrics*, 110(1):371–383.

Harzing, A.-W., Alakangas, S., and Adams, D. (2014). hia: an individual annual h-index to accommodate disciplinary and career length differences. *Scientometrics*, 99(3):811–821.

Harzing, A.-W. and Mijnhardt, W. (2015). Proof over promise: towards a more inclusive ranking of dutch academics in economics & business. *Scientometrics*, 102(1):727–749.

Harzing, A.-W. and van der Wal, R. (2009). A google scholar h-index for journals: An alternative metric to measure journal impact in economics and business. *J. Am. Soc. Inf. Sci. Technol.*, 60(1):41–46.

Haustein, S., Costas, R., and Larivière, V. (2015a). Characterizing social media metrics of scholarly papers: The effect of document properties and collaboration patterns. *PLoS ONE*, 10(3):1–21.

Haustein, S., Sugimoto, C. R., and Larivière, V. (2015b). Guest editorial: Social media metrics in scholarly communication. *CoRR*, abs/1504.01877.

Heer, J., Card, S. K., and Landay, J. (2005). Prefuse: A toolkit for interactive information visualization. In *ACM Human Factors in Computing Systems (CHI)*, pages 421–430.

Henkin, R. and Dykes, J. & Slingsby, A. (2016). Characterizing representation of temporal data visualization. *Poster presented at the VIS 2016*.

Henzinger, M. R., Motwani, R., and Silverstein, C. (2002). Challenges in web search engines. *SIGIR Forum*, 36(2):11–22.

Hernández, M. A. and Stolfo, S. J. (1998). Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem. *Data Min. Knowl. Discov.*, 2(1):9–37.

Herther, N. K. (2011). Scholar citations-google moves into the domain of web of science and scopus. Disponible en: `http://tinyurl.com/hkq7bxa` Visitado el: 03/10/2016.

Hettenhausen, J., Lewis, A., and Mostaghim, S. (2010). Interactive multi-objective particle swarm optimization with heatmap-visualization-based user interface. *Engineering Optimization*, 42(2):119–139.

Hiemstra, D. (2000). A probabilistic justification for using tf-idf term weighting in information retrieval. *International Journal on Digital Libraries*, 3(2):131–139.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569–16572.

Hoang, D. T., Kaur, J., and Menczer, F. (2010). Crowdsourcing Scholarly Data. In *Proc. Web Science Conference: Extending the Frontiers of Society On-Line (WebSci10)*, Raleigh, NC: US.

Huang, J., Ertekin, S., and Giles, C. L. (2006). *Efficient Name Disambiguation for Large-Scale Databases*, pages 536–544. Springer Berlin Heidelberg, Berlin, Heidelberg.

Hug, S. E. and Brandle, M. P. (2017). The coverage of microsoft academic: Analyzing the publication output of a university. *CoRR*, abs/1703.05539.

Hug, S. E., Ochsner, M., and Brandle, M. P. (2016). Citation analysis with microsoft academic. *CoRR*, abs/1609.05354.

Igual Camacho, C. and Díaz Díaz, B. (2008). Producción científica de esclerosis múltiple y su evolución en la última década. *Fisioterapia*, 30(6):262–267.

Jacso, P. (2005). As we may search-comparison of major features of the web of science, scopus, and google scholar citation-based and citation-enhanced databases. *Current Science*, 89(9):1537–1547.

Jacso, P. (2008a). Testing the calculation of a realistic h-index in Google Scholar, Scopus, and Web of Science for F. W. Lancaster. *Library Trends*, 56(4):784–815.

Jacso, P. (2008b). The pros and cons of computing the h-index using Google Scholar. *Online Information Review*, 32(3):437–452.

Jacso, P. (2009). Five-year impact factor data in the journal citation reports. *Online information review*, 33(3):603–614.

Jacso, P. (2012). Google Scholar Author Citation Tracker: is it too little, too late? *Online Information Review*, 36(1):126–141.

Jaro, M. A. (1989). Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84(406):414–420.

Jaro, M. A. (1995). Probabilistic Linkage of Large Public Health Data Files. *Statistics in Medicine*, 14:491–498.

Jeng, W., He, D., and Jiang, J. (2015). User participation in an academic social networking service: A survey of open group users on mendeley. *Journal of the Association for Information Science and Technology*, 66(5):890–904.

Jiang, T. and Gao, H. (2016). Are mendeley's public groups effective aggregators of high-value papers? an analysis based on paper readerships. *IConference 2016*.

Jimenez, S., Becerra, C., Gelbukh, A., and Gonzalez, F. (2009). *Generalized Mongue-Elkan Method for Approximate Text String Comparison*, pages 559–570. Springer Berlin Heidelberg, Berlin, Heidelberg.

Jin, B. (2006). h-index: an evaluation indicator proposed by scientist. *Science Focus*, pages 8–9.

Jin, B. (2007). The AR-index: complementing the h-index. *ISSI Newsletter*, 3(1):6.

Jin, B., Liang, L., Rousseau, R., and Egghe, L. (2007). The R- and AR-indices: Complementing the h-index. *Chinese Science Bulletin*, 52(6):855–863.

John, C. and Joseph, K. (1998). Interactive visualization of serial periodic data. In *ACM Symposium on User Interface Software and Technology, (San Francisco, CA), ACM Press: New York*, pages 29–38.

Kalmar, P. and Freitag, D. (2009). Features for web person disambiguation. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*.

Kang, I. S., Na, S. H., Lee, S., Jung, H., Kim, P., and Sung, W. K. (2009). On co-authorship for author disambiguation. *Information Processing and Management*.

Kaur, J., Hoang, D. T., Sun, X., Possamai, L., JafariAsbagh, M., Patil, S., and Menczer, F. (2012). Scholarometer: A Social Framework for Analyzing Impact across Disciplines. *PLoS ONE*, 7(9).

Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 7(1).

Keim, D. A., Schneidewind, J., and Sips, M. (2004). Circleview: A new approach for visualizing time-related multidimensional data sets. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '04, pages 179–182, New York, NY, USA. ACM.

Kern, R., Zechner, M., and Granitzer, M. (2011). Model selection strategies for author disambiguation. *IEEE Computer Society: 8th International Workshop on Text-based Information Retrieval in Procceedings of 22th International Conference on Database and Expert Systems Applications (DEXA 11)*, pages 155–160.

Khabsa, M. and Giles, C. L. (2014). The number of scholarly documents on the public web. *PLoS One*, 9(5).

Knuth, D. E. (1998). *Art of Computer Programming, Volume 3: Sorting and Searching (2nd Edition)*. Addison-Wesley Professional, 2 edition.

Kosara, R., Hauser, H., and Gresh, D. (2003). An interaction view on information visualization. In *EUROGRAPHICS 2003 State-of-the-Art Re-ports*, pages 123–137.

Kraker, P., Lex, E., Gorraiz, J., Gumpenberger, C., and Peters, I. (2015). Research data explored II: the anatomy and reception of figshare. *CoRR*, abs/1503.01298.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22:49–86.

Labbé, C. (2010). Ike Antkare, one of the great stars in the scientific firmament. *ISSI Newsletter*, 6(2):48–52.

Lacey, M. (2017). Statistical topics. Disponible en: `http://www.stat.yale.edu/Courses/1997-98/101/scatter.htm`, Visitado el: 04/05/2017.

Lee, D., On, B.-W., Kang, J., and Park, S. (2005). Effective and scalable solutions for mixed and split citation problems in digital libraries. In *IQIS '05: Proceedings of the 2nd international workshop on Information quality in information systems*, pages 69–76, New York, NY, USA. ACM Press.

Lee, M. L., Ling, T. W., and Low, W. L. (2000). Intelliclean: A knowledge-based intelligent data cleaner. In *6th International Conference on Knowledge Discovery and Data Mining*, pages 290–294.

Lerchenmueller, M. J. and Sorenson, O. (2016). Author disambiguation in pubmed: Evidence on the precision and recall of author-ity among nih-funded scientists. *PLoS ONE*, 11(7):1–13.

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707+.

Lewis, D. D. (1998). *Naive (Bayes) at forty: The independence assumption in information retrieval*, pages 4–15. Springer Berlin Heidelberg, Berlin, Heidelberg.

Leydesdorff, L., de Moya-Anegón, F., and Guerrero-Bote, V. P. (2010). Journal maps on the basis of scopus data: A comparison with the journal citation reports of the isi. *J. Am. Soc. Inf. Sci. Technol.*, 61(2):352–369.

Li, G.-C., Lai, R., D´Amour, A., Doolin, D. M., Sun, Y., Torvik, V. I., Yu, A. Z., and Fleming, L. (2014). Disambiguation and co-authorship networks of the u.s. patent inventor database (1975–2010). *Research Policy*, 43(6):941 – 955.

Liu, W., Islamaj Dogan, R., Kim, S., Comeau, D. C., Kim, W., Yeganova, L., Lu, Z., and Wilbur, W. J. (2014a). Author name disambiguation for pubmed. *Journal of the Association for Information Science and Technology*, 65(4):765–781.

Liu, X., Yu, Y., Guo, C., and Sun, Y. (2014b). Meta-path-based ranking with pseudo relevance feedback on heterogeneous graph for citation recommendation. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 121–130, New York, NY, USA. ACM.

Liu, Y., Li, W., Huang, Z., and Fang, Q. (2015). A fast method based on multiple clustering for name disambiguation in bibliographic citations. *Journal of the Association for Information Science & Technology*, 66(3):634–644.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences.*, 16(12):317–323.

Low, W. L., Lee, M. L., and Ling, T. W. (2001). A knowledge-based approach for duplicate elimination in data cleaning. *Inf. Syst.*, 26:585–606.

Mainardi, C. F. and de Morán-Suárez, M. A. (2011). La responsabilidad social corporativa (rsc) en las bases de datos scopus y wos (estudio bibliomÉtrico). *EDICIC*, 1(4):141–160.

Martín-Martín, A., Ayllón, J. M., Orduña-Malea, E., and López-Cózar, E. D. (2016a). 2016 google scholar metrics released: a matter of languages... and something else. *CoRR*, abs/1607.06260.

Martín-Martín, A., Ayllón, J. M., Orduña-Malea, E., and López-Cózar, E. D. (2016b). Proceedings scholar metrics: H index of proceedings on computer science, electrical & electronic engineering, and communications according to google scholar metrics (2010-2014). *CoRR*, abs/1606.05341.

Martín-Martín, A., Orduña-Malea, E., Ayllon, J. M., and López-Cózar, E. D. (2016c). The counting house: measuring those who count. presence of bibliometrics, scientometrics, informetrics, webometrics and altmetrics in the google scholar citations, researcherid, researchgate, mendeley & twitter. *CoRR*, abs/1602.02412.

Martín-Martín, A., Orduña-Malea, E., Ayllón, J. M., and López-Cózar, E. D. (2016d). A two-sided academic landscape: portrait of highly-cited documents in google scholar (1950-2013). *CoRR*, abs/1607.02861.

Martín, S. G. (2013). El DOI en las revistas científicas del portal SciELO. *Palabra clave*, 3(1):12–29.

Meddings, K. (2017). Funder registry. Disponible en: `https://www.crossref.org/services/funder-registry/`, Visitado el: 17/02/2017.

Medrano, J. F., Figuerola, C. G., and Alonso Berrocal, J. L. (2012a). Desambiguación de publicaciones científicas, un enfoque práctico. In *I Seminario Hispano Brasileño Investigación, Documentación y Sociedad.*

Medrano, J. F., Figuerola, C. G., and Alonso Berrocal, J. L. (2012b). Repositorios digitales en españa y calidad de metadatos. *Scire. Representación Y Organización Del Conocimiento*, 18(2):109–121.

Meeks, E. (2015). *D3.js in Action.* Manning Publications Co.

Meho, L. I. and Yang, K. (2006). Multi-faceted approach to citation-based quality assessment for knowledge management. *World library and information congress: 72nd IFLA General conference and council.*

Meho, L. I. and Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of science versus scopus and google scholar. *J. Am. Soc. Inf. Sci.*, 58(13):2105–2125.

Melville, P., Yang, S. M., Saar-tsechansky, M., and Mooney, R. (2005). Active learning for probability estimation using Jensen-Shannon divergence. In *Proceedings of the European Conference on Machine Learning (ECML-05*, pages 268–279.

Merton, R. K. (1977). *La sociología de la ciencia, 2: investigaciones teóricas y empíricas*. Alianza Editorial.

Miguel, S. and Solana, V. H. (2010). Visibilidad de las revistas latinoamericanas de bibliotecología y ciencia de la información a través de Google Scholar. *Ciência da Informação*, 39(2).

Ministerio de Ciencia Tecnología e Innovación Productiva and Consejo Interinstitucional de Ciencia y Tecnología (2013). Directrices snrd. directrices para proveedores de contenido del sistema nacional de repositorios digitales ministerio de ciencia, tecnología e innovación productiva. Disponible en: `http://repositorios.mincyt.gob.ar/pdfs/Directrices_SNRD_2013.pdf`, Visitado el: 30/04/2017.

Méndez, E. (2015). Cultura abierta: conocimiento compartido. *Anuario ThinkEPI*, 9:126–131.

Mockapetris, P. V. (1983). Domain Names: Concepts and Facilities. STD 13, RFC 1034.

Mockapetris, P. V. (1987). Domain Names - Implementation and Specification. STD 13, RFC 1035.

Moed, H. F. (2006). *Citation Analysis in Research Evaluation*, volume 9. Springer Science & Business Media.

Moed, H. F. (2009). Measuring contextual citation impact of scientific journals. *CoRR*, abs/0911.2632.

Moed, H. F., Bar-Ilan, J., and Halevi, G. (2016). A new methodology for comparing google scholar and scopus. *Journal of Informetrics*, 10(2):533 – 551.

Momeni, F. and Mayr, P. (2016). Evaluating co-authorship networks in author name disambiguation for common names. *CoRR*, abs/1606.03857.

Monge, A. E. (2001). An adaptive and efficient algorithm for detecting approximately duplicate database records. *International Journal on Information Systems Special Issue on Data Extraction, Cleaning, and Reconciliation*.

Monge, A. E. and Elkan, C. P. (1996). The field matching problem: Algorithms and applications. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 267–270.

Monge, A. E. and Elkan, C. P. (1997). An Efficient Domain-Independent Algorithm for Detecting Approximately Duplicate Database Records. In *SIGMOD workshop on data mining and knowledge discovery*.

Mongeon, P. and Paul-Hus, A. (2016). The journal coverage of web of science and scopus: a comparative analysis. *Scientometrics*, 106(1):213–228.

Munzner, T. (2000). *Interactive Visualization of Large Graphs and Networks*. Phd's thesis, Department of Computer Science, Stanford University.

Murray, S. (2013). *Interactive Data Visualization for the Web*. O'Reilly.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453.

Niyazov, Y., Vogel, C., Price, R., Lund, B., Judd, D., Akil, A., Mortonson, M., Schwartzman, J., and Shron, M. (2016). Open access meets discoverability: Citations to articles posted to academia.edu. *PLoS ONE*, 11(2):1–23.

On, B. W., Lee, D., Kang, J., and Mitra, P. (2005). Comparative study of name disambiguation problem using a scalable blocking-based framework. In *JCDL '05: Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pages 344–353, New York, NY, USA. ACM.

OpenAIRE (2013). Openaire guidelines. Disponible en: `https://guidelines.openaire.eu/en/latest/`, Visitado el: 30/04/2017.

OpenContent.org (2016). Defining the .ºpenïn Open Content and Open Educational Resources. Disponible en: https://www.opencontent.org/definition/ Visitado el: 23/11/2016.

OpenDefinition (2016). Definición de conocimiento abierto. Disponible en: http://opendefinition.org/od/2.0/es/ Visitado el: 23/11/2016.

ORCID (2015). About orcid. Disponible en: http://orcid.org/content/about-orcid Visitado el: 06/11/2016.

Orduña-Malea, E., Ayllón, J., Martín-Martín, A., and Delgado López-Cózar, E. (2014). Empirical evidences in citation-based search engines: is microsoft academic search dead? *ArXiv e-prints*.

Orduña-Malea, E., Martín-Martín, A., Ayllón, J. M., and Delgado López-Cózar, E. (2016a). *La revolución Google Scholar. Destapando la caja de Pandora académica*. Editorial Universidad de Granada, UNE. Unión de Editoriales Universitarias Españolas.

Orduña-Malea, E., Martín-Martín, A., and Delgado López-Cózar, E. (2016b). Researchgate como fuente de evaluación científica: desvelando sus aplicaciones bibliométricas. *El profesional de la información*, 25(2):303–310.

Ortega, J. L. and Aguillo, I. F. (2014). Microsoft academic search and google scholar citations: Comparative analysis of author profiles. *Journal of the Association for Information Science and Technology*, 65(6):1149–1156.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project.

Pascual Cid, V. (2010). *Visual Exploration of Web Spaces*. PhD thesis, Universitat Pompeu Fabra. Departament de Tecnologies de la Informació i les Comunicacions.

Pasula, H., Marthi, B., Milch, B., Russell, S., and Shpitser, I. (2003). Identity uncertainty and citation matching. In *In NIPS*. MIT Press.

Pauly, D. and Stergiou, K. L. (2005). Equivalence of results from two citation analyses: Thompson ISI's citation index and Google's Scholar service. *Ethics in Science and Environmental Politics*, 154(3):33–35.

Pavlech, L. L. (2016). Data citation index. *Journal of the Medical Library Association: JMLA*, 104(1):88–90.

Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. (2007). Object retrieval with large vocabularies and fast spatial matching. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8.

Philips, L. (1990). Hanging on the metaphone. *Computer Language*, 7(12):39–43.

Philips, L. (2000). The double-metaphone search algorithm. *C/C++ User's Journal*, 18(6).

Pieper, D. and Summann, F. (2015). 10 years of Bielefeld Academic Search Engine (BASE): Looking at the past and future of the world wide repository landscape from a service providers perspective. In *Presented at the OR2015. 10th International Conference on Open Repositories, Indianapolis.*

Priem, J. (2014). Altmetrics. In Cronin, B. and Sugimoto, C. R., editors, *Beyond bibliometrics: harnessing multidimensional indicators of performance*, pages 263–287. MIT Press.

Priem, J., Taraborelli, D., Groth, P., and Neylon, C. (2010). Altmetrics: A manifesto, (v.1.0). Disponible en: `http://altmetrics.org/manifesto` Visitado el: 03/10/2016.

Prinz, W. (2006). The Graph Visualization System (GVS): A Flexible Java Framework for Graph Drawing. Master's thesis, Graz University of Technology.

Radicchi, F., Fortunato, S., and Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, 105(45).

Reilly, S. (2013). Handle System or Digital Object Identifiers. Disponible en: `https://standards.data.gov.uk/proposal/handle-system-or-digital-object-identifiers`, Visitado el: 16/02/2017.

Reimer, T. (2015). Your name is not good enough: introducing the orcid researcher identifier at imperial college london. *Insights*, 28(3):76–82.

Research, M. (2016). Microsoft academic - faq. Disponible en: https://microsoftacademic.uservoice.com/knowledgebase/articles/838965-microsoft-academic-faq Visitado el: 03/10/2016.

ResearchGate (2016). 8 out of 8 million. Disponible en: https://www.researchgate.net/blog/post/8-out-of-8-million Visitado el: 07/10/2016.

Restrepo Arango, C. and Urbizagástegui Alvarado, R. (2010). La productividad de los autores en la ciencia de la información colombiana. *Ci. Inf.*, 39(3):09 – 22.

Reuters, T. (2016a). Researcherid. Disponible en: http://wokinfo.com/researcherid/ Visitado el: 07/10/2016.

Reuters, T. (2016b). Subject area terms. Disponible en: http://images.webofknowledge.com/WOKRS56B5/help/WOS/hp_subject_area_terms_easca.html Visitado el: 02/10/2016.

Reuters, T. (2016c). Web of science. Disponible en: http://ipscience.thomsonreuters.com/product/web-of-science/?utm_source=false&utm_medium=false&utm_campaign=false Visitado el: 02/10/2016.

Rico-Sulayes, A. (2015). An evaluation measurement in automatic text classification for authorship attribution. *Ingenio Magno*, 6(2):62–74.

Ristad, E. S. and Yianilos, P. N. (1998). Learning string-edit distance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(5):522–532.

Rizkallah, J. and Sin, D. D. (2010). Integrative approach to quality assessment of medical journals using impact factor, eigenfactor, and article influence scores. *PLoS ONE*, 5(4):e10204.

Robinson-García, N., Torres-Salinas, D., Zahedi, Z., and Costas, R. (2014). New data, new possibilities: Exploring the insides of altmetric.com. *El profesional de la información*, 23(4).

Robinson-Garcia, N., Jiménez-Contreras, E., and Torres-Salinas, D. (2015). Analyzing data citation practices using the data citation index. *Journal of the Association for Information Science and Technology*, 2.

Roemer, R. C. and Borchardt, R. (2012). From bibliometrics to altmetrics. a changing scholarly landscape. *College & Research Libraries News*, 73(10):596–600.

Rousseau, R. (2006). New developments related to the hirsch index. *Science Focus*, 1(4):23–25.

Rousseau, R. (2008). Reflections on recent developments of the h-index and h-type indices. *In H. Kretschmer & F. Havemann (Eds.), Proceedings of WIS 2008. Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting.*

Russell, R. (1918). Index (soundex patent). Technical report, US Patent No. 1,261,167, pp.1:4. 1918.

Sancho, R. (1990). Indicadores bibliométricos utilizados en la evluación de la ciencia y la tecnología. revisión bibliográfica. *Revista española de documentación científica*, 13(4):842–865.

Sankey, H. R. (1898). Introductory note on the thermal efficiency of steam-engines. report of the committee appointed on the 31st march, 1896, to consider and report to the council upon the subject of the definition of a standard or standards of thermal efficiency for steam-engines: With an introductory note. *Minutes of the Proceedings of the Institution of Civil Engineers*, 134:278–283.

Santa, S. and Herrero-Solana, V. (2010). Cobertura de la ciencia de américa latina y el caribe en scopus vs web of science. *Investigación bibliotecológica*, 24(52).

Schmidt, M. (2008). The sankey diagram in energy and material flow management. *Journal of industrial ecology*, 12(1):82–94.

Schreiber, M. (2008a). The influence of self-citation corrections on egghe's g index. *Scientometrics*, 76(1):187–200.

Schreiber, M. (2008b). To share the fame in a fair way, hm modifies h for multi-authored manuscripts. *New J. Phys.*, 10(4):040201+.

Schreiber, M. (2010a). A new family of old hirsch index variants. *Journal of Informetrics*, 4(4):647–651.

Schreiber, M. (2010b). Revisiting the g-index: The average number of citations in the g-core. *J. Am. Soc. Inf. Sci. Technol.*, 61(1):169–174.

Seol, J.-W., Kim, K.-Y., Park, J.-H., Lee, H.-J., Yoon, J.-S., You, B.-J., and Lee, S.-H. (2016). Expanding co-author network for author disambiguation in scholarly data. *Advanced Science and Technology Letters*, 122:186–191.

Serva, M. and Petroni, F. (2007). Indo-European languages tree by Levenshtein distance. *EPL (Europhysics Letters)*.

Services, M. C. (2017). Paper entity. documentation. Disponible en: `https://www.microsoft.com/cognitive-services/en-us/Academic-Knowledge-API/documentation/EntityAttributes/PaperEntity`, Visitado el: 02/03/2017.

Shneiderman, B. (1998). *Designing the User Interface*. Addison-Wesley, Reading, MA., 3 edition.

Sidiropoulos, A., Katsaros, D., and Manolopoulos, Y. (2007). Generalized hirsch h-index for disclosing latent facts in citation networks. *Scientometrics*, 72(2):253–280.

Silva, S. F. and Catarci, T. (2000). Visualization of linear time-oriented data: A survey. In *Proceedings of the First International Conference on Web Information Systems Engineering (WISE'00)-Volume 1 - Volume 1*, WISE '00, pages 310–, Washington, DC, USA. IEEE Computer Society.

Simonton, D. K. (1999). *Creativity and genius. Handbook of personality theory and research*. New York: Guilford Press. L. A. Pervin y O. John (eds.).

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J. P., and Wang, K. (2015). An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 243–246, New York, NY, USA. ACM.

Snae, C. (2007). A comparison and analysis of name matching algorithms. *Engineering and Technology*, 21(January):252–257.

SocietyZone (2016). Managing and discovering academic references: Citeulike and its social bookmarking service. Disponible en: `http://tinyurl.com/z4jldnp` Visitado el: 09/10/2016.

Soon, W. M., Ng, H. T., and Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Comput. Linguist.*, 27(4):521–544.

Springer (2015). Explore the impact of your books! Disponible en: `http://www.springer.com/gp/authors-editors/book-authors-editors/bookmetrix` Visitado el: 06/11/2016.

Stefano, D. D., Fuccella, V., Vitale, M. P., and Zaccarin, S. (2013). The use of different data sources in the analysis of co-authorship networks and scientific performance. *Social Networks*, 35(3):370–381.

Stonebraker, J. S., Gil, E., Kirkwood, C. W., and Handfield, R. B. (2012). Impact factor as a metric to assess journals where om research is published. *Journal of Operations Management*, 30(1):24 – 43.

Suber, P. (2008). Gratis and libre open access. SPARC Open Access Newsletter, August 2008 issue. Disponible en: `http://sparcopen.org/our-work/gratis-and-libre-open-access/` Visitado el: 23/11/2016.

Suber, P. (2012). *Open Access (MIT Press Essential Knowledge)*. The MIT Press.

Sun, S., Lannom, L., and B., B. (2003). Handle System Overview, RFC 3650. Disponible en: `https://www.rfc-editor.org/info/rfc3650`.

Sun, X., Kaur, J., Possamai, L., and Menczer, F. (2013). Ambiguous author query detection using crowdsourced digital library annotations. *Information Processing and Management*, 49(2):454–464.

Tang, L. and Walsh, J. P. (2010). Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps. *Scientometrics*, 84(3):763–784.

Tata, S. and Patel, J. M. (2007). Estimating the selectivity of *tf-idf* based cosine similarity predicates. *SIGMOD Rec.*, 36:75–80.

Tejada, S., Knoblock, C. A., and Minton, S. (2002). Learning domain-independent string transformation weights for high accuracy object identification. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 350–359, New York, NY, USA. ACM.

Teller, S. (2013). *Data Visualization with d3.js*. Packt Publishing.

The University of Chicago Library (2016). Author and research identifiers. Disponible en: http://guides.lib.uchicago.edu/c.php?g=298332&p=1989825 Visitado el: 06/11/2016.

The University of the Sunshine Coast Library (2016). Researcher identifiers and your online research profile. Disponible en: http://libguides.usc.edu.au/c.php?g=508471&p=3478537 Visitado el: 06/11/2016.

Thelwall, M. and Kousha, K. (2014). Academia.edu: Social network or academic network? *Association for Information Science and Technology*, 65(4):721–731.

Thelwall, M. and Kousha, K. (2015). Researchgate: Disseminating, communicating, and measuring scholarship? *Journal of the Association for Information Science & Technology*, 66(5):876–889.

Thelwall, M. and Kousha, K. (2016). Figshare: a universal repository for academic resource sharing? *Online Information Review*, 40(3):333–346.

Thelwall, M., Vaughan, L., and Björneborn, L. (2005). Webometrics. *Annual Review of Information Science and Technology*, 39(1):81–135.

Thomas, W. J., Chen, B., and Clement, G. (2015). Orcid identifiers: Planned and potential uses by associations, publishers, and librarians. *The Serials Librarian*, 68(1-4):332–341.

Thor, A. and Bornmann, L. (2011). The calculation of the single publication h index and related performance measures A web application based on Google Scholar data. *ONLINE INFORMATION REVIEW*, 35(2):291–300.

Tol, R. S. J. (2009). The h-index and its alternatives: An application to the 100 most prolific economists. *Scientometrics*, 80(2):317–324.

Tomlin, P. (2009). A matter of discipline: Open access, the humanities, and art history. *Canadian Journal of Higher Education*, 39(3):49–69. Disponible en: http://ojs.library.ubc.ca/index.php/cjhe/ Visitado el: 23/11/2016.

Torres-Salinas, D., Cabezas-Clavijo, A., and Jiménez-Contreras, E. (2013). Altmetrics: New indicators for scientific communication in web 2.0. *CoRR*, abs/1306.6595.

Torres-Salinas, D. and Jiménez-Contreras, E. (2010). Introducción y estudio comparativo de los nuevos indicadores de citación sobre revistas científicas en journal citation reports y scopus. *El profesional de la información*, 19(2):201–207.

Torres-Salinas, D., Jiménez-Contreras, E., and Robinson-García, N. (2014a). How many citations are there in the data citation index? *19th International Conference on Science and Technology Indicators*.

Torres-Salinas, D., Martín-Martín, A., and Fuente-Gutiérrez, E. (2014b). Analysis of the coverage of the data citation index thomson reuters: disciplines, document types and repositories. *Revista Española de Documentación Científica*, 37(1).

Torres-Salinas, D. and Milanés-Guisado, Y. (2014). Presencia en redes sociales y altmétricas de los principales autores de la revista el profesional de la información. *El profesional de la información*, 23(4):367–372.

Torres-Salinas, D., Ruiz-Pérez, R., and Delgado-López-Cózar, E. (2009). Google Scholar como herramienta para la evaluación científica. *El Profesional de la Información*, 18(5):501–510.

Tran, H. N., Huynh, T., and Do, T. (2014). *Author Name Disambiguation by Using Deep Neural Network*, pages 123–132. Springer International Publishing, Cham.

Tria, F., Caglioti, E., Loreto, V., and Pagnani, A. (2010). A stochastic local search approach to language tree reconstruction. *Diachronica*, 27(2):341–358.

Triesman, A. and Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review*, 95(1):15–48.

Tufte, E. R. (1986). *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, USA.

University of Tasmania (2016). Research identity: Scopus author id. Disponible en: `http://utas.libguides.com/c.php?g=498459&p=3411328` Visitado el: 06/11/2016.

Urbizagástegui Alvarado, R. (2005). La productividad científica de los autores: Un modelo de aplicación de la ley de lotka por el método del poder inverso generalizado. *Información, cultura y sociedad*, 1(12):51–73.

Valarakos, A. G., Valarakos, R. G., Paliouras, G., Karkaletsis, V., and Vouros, G. (2004). A name-matching algorithm for supporting ontology enrichment. In *Proceedings of SETN'04, 3rd Hellenic Conference on Artificial Intelligence*, pages 381–389. Springer Verlag.

Valderrama-Zurián, J.-C., Aguilar-Moya, R., Melero-Fuentes, D., and Aleixandre-Benavent, R. (2015). A systematic analysis of duplicate records in scopus. *Journal of Informetrics*, 9(3):570 – 576.

Van de Sompel, H. and Lagoze, C. (2000). The Santa Fe Convention of the Open Archives Initiative. *D-Lib Magazine*, 6(2).

Van Noorden, R. (2010). Metrics: A profusion of measures. *Nature*, 465(7300):864–866.

Vanclay, J. K. (2012). Impact factor: outdated artefact or stepping-stone to journal certification? *Scientometrics*, 92(2):211–238.

Walters, W. H. (2011). Comparative recall and precision of simple and expert searches in google scholar and eight other databases. *Portal: Libraries and the Academy*, 11(4):971–1006.

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

Wang, Q. and Waltman, L. (2016). Large-scale analysis of the accuracy of the journal classification systems of web of science and scopus. *Journal of Informetrics*, 10(2):347 – 364.

Ware, C. (2004). *Information Visualization - Perception for Design.* Morgan-Kaufmann, 2 edition.

Ware, C. (2008). *Visual Thinking for Design.* Morgan Kaufman/Elsevier.

Washio, T., Inokuchi, A., Suzuki, E., and Ting, K. M., editors (2008). *PAKDD'08: Proceedings of the 12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, Berlin, Heidelberg. Springer-Verlag.

Wheeler, L. (2015). Digital Science Launches GRID, a New, Global, Open Database Offering Unique Information on Research Organisations. Disponible en: `https://tinyurl.com/zvqo8by`, Visitado el: 16/02/2017.

Wichmann, S. and Holman, E. W. (2009). Population size and rates of language change. *Human Biology*, 81(2):259–274.

Wiley, D. (2014). The access compromise and the 5th r. Disponible en: `http://opencontent.org/blog/archives/3221` Visitado el: 23/11/2016.

Winkler, W. E. (1999). The State of Record Linkage and Current Research Problems. Technical report, Statistical Research Division, U.S. Census Bureau.

Xiao, C., Wang, W., Lin, X., Yu, J. X., and Wang, G. (2011). Efficient similarity joins for near-duplicate detection. *ACM Transactions on Database Systems*, 36.

Yang, H. and Callan, J. (2006). Near-duplicate detection by instance-level constrained clustering. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 421–428, New York, NY, USA. ACM.

Yang, K. and Meho, L. I. (2006). Citation Analysis: A Comparison of Google Scholar, Scopus, and Web of Science. *Proceedings of the American Society for Information Science and Technology*, 43(1):1–15.

Yu, M.-C., Wu, Y.-C. J., Alhalabi, W., Kao, H.-Y., and Wu, W.-H. (2016). ResearchGate: An effective altmetric indicator for active researchers? *Computers in Human Behavior*, 55:1001–1006.

Zeng, Y., Yao, Y., and Zhong, N. (2009). DBLP-SSE: A DBLP Search Support Engine. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '09, pages 626–630, Washington, DC, USA. IEEE Computer Society.

Zhang, B., Dundar, M., and Al Hasan, M. (2016). Bayesian non-exhaustive classification A case study: Online name disambiguation using temporal record streams. *CoRR*, abs/1607.05746.

Zhang, C.-T. (2009). The e-index, complementing the h-index for excess citations. *PLoS ONE*, 5(5).

Zhu, N. Q. (2013). *Data Visualization with D3.js Cookbook*. Packt Publishing.