UNIVERSIDAD METROPOLITANA DE CIENCIAS DE LA EDUCACIÓN

FACULTAD DE CIENCIAS BÁSICAS

DEPARTAMENTO DE QUÍMICA

ANÁLISIS BIBLIOMÉTRICO DE BASES DE DATOS CIENTÍFICAS EN LÍNEA Y UN NUEVO FACTOR DE IMPACTO INTERNO COMO DERIVACIÓN DEL ÍNDICE H

Tesis para optar al título de Profesor de Química con mención
Educación en Tecnología

AUTOR: FREDDY AGUSTÍN CARRASCO HERNÁNDEZ

PROFESOR GUÍA: DR. JORGE RODRÍGUEZ BECERRA

La presente Tesis contó con el apoyo financiero del proyecto FONDECYT – 1221942

SANTIAGO DE CHILE, NOVIEMBRE DE 2022

IDENTIFICACIÓN DE TESIS/INVESTIGACIÓN

Título de la tesis: Análisis bibliométrico de bases de datos científicas en línea y un nuevo factor de impacto interno como derivación del índice H

Fecha: Noviembre 2022

Facultad: Ciencias Básicas

Departamento: Química

Carrera: Licenciatura en Educación en Química y Pedagogía en Química con mención Educación en Tecnología

Título y/o grado: Profesor de Química con Mención Educación Tecnológica

Profesor guía: Prof. Dr. Jorge Rodríguez Becerra

AUTORIZACIÓN: Autorizo a través de este documento, la reproducción total o parcial de este trabajo de investigación para fines académicos, su alojamiento y publicación en el repositorio institucional SIBUMCE del Sistema de Biblioteca UMCE.

Freddy Agustín Carrasco Hernández

Santiago de Chile, noviembre 2022

Agradecimientos:

A mi familia. Mis padres Luis Carrasco y Gloria Hernández, quienes siempre me dieron todo, incluso más de lo que podían, siempre motivándome a estudiar aun cuando ya no me quedaban energías, de pequeño me llenaron de libros y también de amor. Mi hermano Diego Carrasco, mi compadre, en quien siempre he confiado, mi gran amigo, quien me ha enseñado de cables y computadores, pero mucho más sobre cómo levantarse una y otra vez. A Nicole Muñoz, con quien por años compartí mi vida, sueños y alegrías, y seguimos maravillándonos de nuestro retoño, Vicente Carrasco, mi soldado, proletario y amigo, quien me muestra día a día lo hermosa que es la vida.

A todos los profesores que ayudaron en mi formación, quienes me hicieron ver en la educación una forma de ayudar a crear un mundo mejor. A los amigos que hice en los trabajos, sobre todo cuando aguantaban que les dejara más pega.

A mi profesor guía el Dr. Jorge Rodríguez Becerra, quien aún recuerda cómo llegue la primera semana de universidad al laboratorio con los computadores, y me ha recibido más de una vez en su laboratorio, creyendo siempre en mí. Él junto con su esposa, la Dra. Lizethly Cáceres Jensen me han mostrado el increíble mundo de la investigación y el amor por la ciencia. Al Dr.(c) José Hernández Ramos, cuyas tesis de pregrado y posgrado han sido muy interesantes y motivantes para seguir este camino de ciencia, computación y educación. A todo el grupo PACHEM Lab, mis compañeros Mauricio, Manuel, Valeria, y Nayareth, con quienes nos apoyamos todos estos meses de ardua labor, compartiendo llantos, pizzas y risas.

Por último, al Fondo Nacional de Desarrollo Científico y Tecnológico, FONDECYT, por su ayuda económica en la realización de este trabajo.

Tabla de contenido

Listado de tablas

Listado de gráficos

Listado de figuras

1. Resumen

El objetivo de esta investigación se centra en caracterizar de forma general la literatura científica del área de las bases de datos web vinculadas a la computación científica mediante un análisis bibliométrico. Este análisis partió con un total de 2565 registros bibliográficos desde *Web of Science*, correspondientes a la literatura de los últimos diez años. A este primer *dataset* se le realizó un acucioso descarte de duplicados mediante EndNote X9 y Microsoft Excel en sus versiones 2016, 2019 y 365, encontrando el segundo duplicados que no encontró el primero. El *dataset* sin duplicados se sometió a un barrido por impacto, para lo cual se generó un nuevo índice de impacto denominado F_H, derivado del índice H, el que se aplicó para seleccionar las revistas de mayor impacto en el tema tuviesen o no trayectoria en él. Como resultado se logra compilar un *dataset* con las 1951 publicaciones referentes a bases de datos web vinculadas a la computación científica de mayor impacto. El análisis bibliométrico realizado logra identificar a un grupo de revistas con mayor influencia en el campo, encabezado por *Nucleic Acids Research*, seguido por *Database: The Journal of Biological Databases and Curation, PLOS ONE* y *Bioinformatics*. El *dataset* producto de esta investigación constituye un insumo de gran relevancia para posteriores análisis en el marco de análisis bibliométricos y revisiones sistemáticas.

2. Glosario

Siglas, abreviaturas y términos

WoS:	*Web of Science*
WoS CC:	*Web of Science Core Collection*
SCI-Expanded:	*Science Citation Index Expanded*
SSCI:	*Social Science Citation Index*
A&HCI:	*Arts & Humanities Citation Index*
DOI:	Identificador de objetos digitales
H:	Índice H
G:	Índice G
m:	Cociente *m*
TC:	Total de citas
TCA:	Total de citas anualizado
LC:	Citas locales
F_H:	Índice F_H
DBMS:	Sistema de gestión de bases de datos
NCEI:	*National Centers for Environmental Information*
NP:	*Natural products*
MS:	Espectrómetro de masas
NMR:	Resonancia magnético nuclear
BDB:	Base de datos bibliométrica
Deduplicación:	Descarte de duplicados

3. Introducción

3.1. Bases de datos

3.1.1. Un primer acercamiento.

Recordar los números de emergencia 131, 132 y 133, para ambulancias, bomberos y policía respectivamente, es de vital importancia, por ello se les enseña a los niños ya en preescolar, para que ante la necesidad de dar aviso de una emergencia puedan llamar rápidamente a estas instituciones. Posterior a ello, se les suele enseñar los números telefónicos de sus padres, los que ya no son de tres cifras si no que, de nueve, y cada nuevo número que se deba memorizar tendrá también nueve cifras.

Esta tarea de recordar la gran cantidad de números telefónicos, hace algunos años atrás era resuelta con dos grandes libros: cada hogar con conexión telefónica cableada recibía un ejemplar de las Páginas Amarillas, donde acudía casi por defecto a buscar el número telefónico de cualquier comercio o servicio de su zona, todo en un gran libro amarillo; este solía estar acompañado de otro libro gordo y pesado como el anterior pero en hojas blancas, en el que se podía encontrar el número telefónico de cualquier persona, con nombre y apellido, algo que hoy nos parecería impensable por la privacidad de los datos personales. Fuera de cualquier cuestionamiento ético, funcionaba. Estas dos grandes agendas telefónicas son un ejemplo de bases de datos común para todos los mayores de 25.

Hoy en día, cada usuario de un teléfono móvil busca en su agenda telefónica —la que ahora está dentro de su teléfono, en una tarjeta de memoria— donde no pesa cerca del kilogramo como antes, de hecho, ahora su peso se mide en bytes. Para comercios o servicios, como buscar una pastelería o un gásfiter, ya no acude a un pesado libro, pues su teléfono es inteligente —*smartphone*— y con él puede buscarlos en Google, el que no solo tiene una gran lista de ellos, si no que le muestra los más cercanos, le indica hasta qué hora atienden e incluso lo puede guiar con un mapa para llegar a la pastelería. Todo es posible gracias al avance de las tecnologías, pero seguimos con la misma idea, bases de datos, más enriquecidas e interconectadas, con innumerables ventajas de las que se manejaban hace décadas atrás.

3.1.2. Cada vez más información. Crece una necesidad.

Un estudio de la *University of California, Berkeley*, logró estimar la cantidad de información producida en solo un año. Considerando medios de almacenamiento como impresiones en papel, cintas, discos duros, y flujos de información oída o vista, transmitida por radio, TV o internet, estimó que el año 2002 en todo el mundo se produjo entre tres y cinco exabytes. Al tomar en cuenta los datos recabados tres años antes[1], para el año 1999 la cifra de producción total habría sido de entre dos y tres exabytes. Su estimación —sobre los límites superiores— indica que la producción de información en los medios fue creciendo a una tasa de más del 30% por año (Lyman, 2003).

Tabla 1. ¿Cuán grande es un exabyte?

Unidad de medida de almacenamiento de datos	**Equivalencia numérica, en potencia y ejemplos**
kilobyte (KB)	1.000 bytes ó 10^3 bytes. 2 kilobytes: Una página escrita a máquina. 100 kilobytes: Una fotografía de baja resolución.
megabyte (MB)	*1.000.000 bytes ó 10^6 bytes.* 1 megabyte: Una novela corta o un antiguo disquete de 3,5. 2 megabytes: Una fotografía de alta resolución. 5 megabytes: Obras completas de Shakespeare. 10 megabytes: Un minuto de audio en alta fidelidad. 100 megabytes: 1 metro de libros apilados. 700 megabytes: Un CD-ROM.
gigabyte (GB)	*1.000.000.000 bytes ó 10^9 bytes.* 1 gigabyte: Una camioneta llena de libros. 20 gigabytes: Una buena colección de las obras de Beethoven. 100 gigabytes: Un piso de una biblioteca de revistas académicas.
terabyte (TB)	*1.000.000.000.000 bytes ó 10^{12} bytes.* 1 terabyte: 50.000 árboles hechos papel e impreso. 2 terabytes: Una biblioteca de investigación académica. 10 terabytes: Las colecciones impresas de la Biblioteca del Congreso de los EEUU. 200 terabytes: Los datos archivados cada mes en la base de datos del NCEI.
petabyte (PB)	*1.000.000.000.000.000 bytes ó 10^{15} bytes* 1 petabyte: Los datos medidos por la EOS entre 1999 y el 2001. 2 petabytes: Todas las bibliotecas de investigación los EEUU. 20 petabytes: Discos duros producidos en 1995. 200 petabytes: Todo el material impreso existente al 2003.
exabyte (EB)	*1.000.000.000.000.000.000 bytes OR 10^{18} bytes* 2 exabytes: Toda la información generada en 1999 (según la estimación mínima). 5 exabytes: Todas las palabras dichas por los seres humanos.

Adaptada de *How Much Information? 2003* por P. Lyman, 2003 (https://groups.ischool.berkeley.edu/archive/how-much-info-2003/execsum.htm). Todos los derechos reservados 2003 por *Regents of the University of California*.

[1] Dicho estudio anterior difiere en algunos aspectos metodológicos, los datos son recalculados en el estudio del 2003.

En la Tabla 1 se pueden dimensionar las enormes cantidades de información que se revelan en el estudio de Berkeley, con ejemplos como la cantidad información contenida en una hoja llena de texto, hasta grandes bases de datos de nivel mundial.

La creciente oleada de información lleva aparejada una creciente necesidad de gestionar estos datos. Así como las bibliotecas que manejan grandes volúmenes de libros de diferentes temas, y se encargan de clasificarlos y ordenarlos para que el usuario/lector encuentre lo que busca, pueda acceder a él, lo pueda leer, y hasta en algunos casos sacarlo de la biblioteca, las bases de datos en general apuntan a lo mismo, pero deben estar preparadas para hacerlo a una mucho mayor escala, se ayudan de *hardware* y *software* adecuados para poder manejar grandes cantidades de información, y deben satisfacer necesidades específicas referentes a los tipos de datos que gestionen, o en otras palabras, tipos de archivos, además de las necesidades asociadas a las diferentes áreas del conocimiento.

3.1.3. Definiciones y antecedentes históricos de las bases de datos.

El término base de datos, fue acuñado por primera vez en 1963, en el simposio *"Development and Management of a Computer-centered Data Base"* (Olle, 2006). Desde ese entonces ha sido definido en reiteradas ocasiones desde la informática. Uno de los grandes referentes en el campo en las últimas décadas es Christopher J. Date, quien las define de una forma simple y acotada como: *"…un conjunto de datos persistentes que es utilizado por los sistemas de aplicación de alguna empresa dada",* ocupando el concepto de empresa de una forma más amplia que lo literal, refiriéndose con ello a cualquier organización independiente de tipo comercial, técnico, científico u otro, pudiendo corresponder ello a un solo individuo con una pequeña base de datos personal o a una gran base de datos compartida por una corporación. (Date, 2001, p. 10)

Una década más tarde, Edgar F. Codd (1970) define el modelo relacional, los fundamentos de la estructura imperante hoy en día para la escritura y programación de las bases de datos, publicando además una serie de reglas para la evaluación de administradores de sistemas de datos relacionales.

A partir de los aportes de Codd, Larry Ellison desarrolló la base de datos Oracle, un sistema de administración de base de datos (DBMS), que destaca por sus transacciones,

estabilidad, escalabilidad y multiplataforma. Hoy, Oracle Corporation, es referente mundial en bases de datos. Desde el sitio web de Oracle se entrega la siguiente definición:

> Una base de datos es una recopilación organizada de información o datos estructurados, que normalmente se almacena de forma electrónica en un sistema informático. Normalmente, una base de datos está controlada por un sistema de gestión de bases de datos. En conjunto, los datos y el DBMS, junto con las aplicaciones asociadas a ellos, reciben el nombre de sistema de bases de datos, abreviado normalmente a solo base de datos.
>
> Los datos de los tipos más comunes de bases de datos en funcionamiento actualmente se suelen utilizar como estructuras de filas y columnas en una serie de tablas para aumentar la eficacia del procesamiento y la consulta de datos. Así, se puede acceder, gestionar, modificar, actualizar, controlar y organizar fácilmente los datos. La mayoría de las bases de datos utilizan un lenguaje de consulta estructurada -SQL- para escribir y consultar datos. (Oracle Corporation, s.f.)

Para complementar, Berrington añade en su definición los conceptos de registros, campos, y toma de decisiones:

> Una base de datos es una colección estructurada de registros o datos que se almacena en una computadora para que pueda ser consultada por un programa para responder consultas. Los registros recuperados a través de consultas se convierten en información que se puede utilizar para tomar decisiones. Una base de datos consta de una o más tablas que contienen registros de valores para campos que pertenecen a los atributos del objeto representado por la tabla. Las bases de datos relacionales contienen varias tablas que están vinculadas por medio de campos clave. Un sistema de gestión de base de datos es

el programa informático que gestiona la base de datos y consulta los datos para producir informes de información. (2017, p. 155)

Sin ahondar en los lenguajes de escritura, como el mencionado por Oracle, se pueden destacar los siguientes puntos: a) Las bases de datos corresponden a "recopilación organizada de información o datos estructurados"; b) Ellas suelen estar asociadas a un sistema que los gestiona (DBMS). Aunque son dos cosas diferentes, es común que se nombre también a este conjunto —datos ordenados y sistema que los gestiona— como base de datos[2].

3.1.3.1. *Dataset*

El término *dataset* (set de datos) no es más que un conjunto de datos estructurados, por lo general correspondientes a una única base de datos de origen. En él, cada fila corresponde a un ítem o registro, mientras que cada columna a una variable o campo.

Los lenguajes en que están escritos -en otras palabras, los formatos de archivo-, hacen que puedan ser leídos por múltiples programas o en algunos casos solo por uno.

3.1.4. Bases de datos para la ciencia.

Una forma para clasificar a los diferentes tipos de bases de datos existentes, es según los lenguajes de programación y otros términos propios de la informática detrás del diseño de las mismas como: Bases de datos SQL o No-SQL, Bases de datos Relacionales o No relacionales, etc. No se ahondará en estas clasificaciones porque se escapa del foco de esta investigación. ¿Cuál es este foco? Las bases de datos enfocadas en las ciencias, o que —aun sin estar diseñadas en dicho marco— pueden resultar útiles para las mismas.

Así como fue demostrado en la investigación de Berkeley (Lyman, 2003), cada año el mundo genera enormes cantidades de información, y las bases de datos son por ello, sistemas sin los cuales estos datos serían imposibles de manejar, o por lo menos, no al veloz ritmo con que avanza hoy.

[2] Ejemplo de ello son las menciones las bases de datos WoS y Scopus. Ambos son sitios web donde es posible realizar búsqueda de artículos científicos y diversa información bibliográfica asociada a ellos, en ese respecto son DBMS, al hacer referencia a un proceso de gestión, pero también se los menciona como el conjunto de información bibliográfica allí almacenada, haciendo referencia a la definición específica de base de datos.

3.1.4.1. *Búsqueda de información y bases de datos bibliométricas*

Google es hoy el principal portal de acceso a la información. Al buscar en Google el término "mesa", el buscador indica obtener "Cerca de 1.590.000.000 resultados (0,69 segundos)", al hacerlo en inglés por "table", "Cerca de 9.280.000.000 resultados (0,65 segundos)"[3]. Tanta información no es manejable por una persona, Google tiene la misión de —incluyendo la búsqueda, gestión, orden y una serie de procesos—mostrar finalmente los mejores resultados, incluso lo hace tomando en cuenta el perfil del usuario. El DBMS de Google logra realizar esta tarea, sobre una gran cantidad de datos, para ofrecerme información, que me ayude a tomar una decisión.

Sin embargo, es muy frecuente que las búsquedas sean mucho más especializadas, y aún con la tecnología del buscador principal de Google —google.com u otras variantes locales como google.cl— la gran cantidad de resultados me lleve a tener que revisar uno por uno los sitios a los cuáles se hace referencia, muchas veces terminando por desistir de la búsqueda, esta es una de las muchas formas en que sufrimos infoxicación[4].

Para búsquedas más eficientes, existen servicios de búsqueda especializados en un área particular. De gran interés en la ciencia son las bases de datos enfocadas en la búsqueda de información científica. Entre estas, las bases de datos bibliométricas son las enfocadas en literatura científica. Se volverá a tocar este tema en al punto **3.2.3**.

3.1.4.2. *Quimio informática y productos naturales. El ejemplo de Sorokina.*

Hace dos años se publicó en la revista *Journal of Cheminformatics* el estudio titulado "*Review on natural products databases: where to find data in 2020*" (Sorokina & Steinbeck, 2020). Los autores resumen:

> Los productos naturales (NP) han sido el centro de atención de la comunidad científica en las últimas décadas y el interés en torno a ellos sigue creciendo sin cesar. Como consecuencia, en los últimos 20 años, hubo una rápida multiplicación de varias bases de

[3] Búsquedas hechas en www.google.cl el 30 de septiembre del 2022.
[4] Término que hace referencia a un exceso de información que nos es imposible manejar Cornella, A. (2000). *Cómo sobrevivir a la infoxicación*. https://web.archive.org/web/20190429043743id_/http://www.infonomia.com/img/pdf/sobrevivir_infoxicacion.pdf.

> datos y colecciones como recursos generalistas o temáticos para la información de NP. En esta revisión, establecemos una descripción completa de estos recursos, y los números son abrumadores: más de 120 bases de datos y colecciones de NP diferentes fueron publicadas y reutilizadas desde el año 2000. 98 de ellas todavía son accesibles de alguna manera y solo 50 son de acceso abierto. Estos últimos incluyen no solo bases de datos sino también grandes colecciones de NP publicadas como material complementario en publicaciones científicas y colecciones que fueron respaldadas en la base de datos ZINC para compuestos disponibles comercialmente. Algunas bases de datos, incluso publicadas hace relativamente poco tiempo, ya no son accesibles, lo que conduce a una pérdida dramática de datos sobre NP. Las fuentes de datos se presentan en este manuscrito, junto con la comparación del contenido de los abiertos. Con esta revisión, también compilamos los compuestos naturales de acceso abierto en un solo *dataset*, una Colección de Productos Naturales Abiertos (COCONUT), que está disponible en Zenodo y contiene estructuras y anotaciones dispersas para más de 400.000 NP no redundantes, lo que la convierte en la mayor colección abierta de NP disponible hasta la fecha. (p. 1)

Las bases de datos analizadas fueron catalogadas según la disponibilidad, pues no todas las bases de datos son abiertas, algunas de ellas restringen el acceso y/o descarga de sus datos. Algunas son comerciales, cobrando altas sumas de dinero para su uso, lo que implica grandes dificultades para su acceso, siendo el costo una barrera para la difusión del conocimiento. Otras por el cambio sí son de acceso abierto (*open-access*).

También entregan una clasificación según el tipo de datos que ofrecen. Solo por nombrar algunas:

a) Bases de datos de metabolitos y compuestos químicos
b) Bases de datos para dereplicación, distinguiendo entre:

    a. Dereplicación por datos de MS
    b. Dereplicación por datos de NMR

c) Bases de datos generalistas de NP

d) Bases de datos temáticas:

e) Bases de datos según las taxonomías de los organismos sintetizantes de los NP:

    a. NP metabolizados por microorganismos
    b. NP metabolizados por plantas

f) Bases de datos según el uso de los NP:

g) NP para uso en medicina tradicional

    a. En Fitoquímica
    b. En Medicina Tradicional China (TCM)
    c. En Medicina Tradicional Africana (ATM)
    d. De NP tipo droga
    e. De nuevos antibióticos

h) Bases de datos según su ubicación geográfica

i) Catálogos industriales

Los autores señalan la gran cantidad de bases de datos encontrada como un problema, y apuntan a que "*This multiplicity of databases comes also from the publishing pressure on scientists, the infamous 'publish or perish'*" [Esta multiplicidad de bases de datos proviene también de la presión editorial sobre los científicos, el infame 'publicar o perecer'] (p. 44), pues incluso en un sub campo de la ciencia como las bases de datos sobre productos naturales está afecta al frenesí por publicar y producir. Incluso muestran como ejemplo una base de datos que después de un año de haberse publicado ya estaba sin mantenimiento.

El crecimiento del conocimiento en un área con el tiempo termina dando paso a la especialización y con ello fragmentación. Para no perder la conexión entre estos fragmentos y así poder generar conocimiento de forma más integradora surgen los estudios que resumen y/o sintetizan a través de revisiones, o la creación de vínculos entre los fragmentos (Kochen, 1963). Este estudio es un ejemplo de ello, pues analiza diferentes fuentes de información, cada una

especializada en una temática dentro de un tema a estudiar, para luego poder ofrecer al lector información que integra dichos conocimientos, logrando dar a conocer un campo de la ciencia en desarrollo, en este caso, las bases de datos de productos naturales.

3.2. Bibliometría

3.2.1. Disciplinas métricas de la información: Bibliometría, Cienciometría e Informetría.

Los términos bibliometría, cienciometría e Informetría, son conceptos relacionados fuertemente, y en esta investigación están presentes los tres. Corresponden a subcampos de la ciencia cuyo origen se remonta a la primera mitad del siglo XX. En primer lugar, ¿cuál es la importancia de estos estudios?

> El análisis y la evaluación de la información y el conocimiento resultante de la actividad científica es un elemento imprescindible para todos los programas de investigación pública, tecnología y desarrollo que se implementan en una sociedad; y es allí donde la Ciencia de la Información brinda una ayuda inestimable, al desarrollar técnicas e instrumentos para medir la producción de conocimiento y su transformación en bienes.
>
> Las disciplinas métricas de la información (Bibliometría, Cienciometría e Informetría) han permitido el desarrollo de indicadores que, al margen de ventajas y limitaciones ampliamente debatidas, y sobre todo cuando son producto de un análisis multifactorial del contexto donde son aplicados, constituyen herramientas clave en la gestión de la política científica y tecnológica, y en los procesos de toma de decisiones estratégicas. (Arencibia-Jorge & de Moya-Anegon, 2008, pp. 1,2)

Las delimitaciones de estos campos no son unánimes para los autores, pero sí, las diferentes definiciones que aportan los autores claves en la materia, permiten comprender cuál es el foco de cada una.

3.2.1.1. *¿Qué es la bibliometría?*

Pritchard (1969) la definió como:

> …la aplicación de las matemáticas y los métodos estadísticos para analizar el curso de la comunicación escrita y el curso de una disciplina. Dicho de otra manera, es la aplicación de tratamientos cuantitativos a las propiedades del discurso escrito y los comportamientos típicos de éste. (p. 348)

Spinak (1996), suma en su definición a la cienciometría, marcando diferencias entre los dos campos. Según él:

> La bibliometría estudia la organización de los sectores científicos y tecnológicos a partir de las fuentes bibliográficas y patentes para identificar a los actores, a sus relaciones y tendencias. La cienciometría, por el contrario, se encarga de la evaluación de la producción científica mediante indicadores numéricos de publicaciones, patentes, etc.. (p. 35)

Camps (2007) dice "*La bibliometría es la ciencia que permite el análisis cuantitativo de la producción científica a través de la literatura, estudiando la naturaleza y el curso de una disciplina científica"*, recalcando su carácter de ciencia, pero para definirla, no dice lo que hace, si no lo que permite, y luego lo que logra con ello, por lo que su definición, aunque no nombra explícitamente a la cienciometría, es complementaria a la de Spinak.

3.2.1.2. *¿Qué es la cienciometría?*

Spinak (1996), se refiere también a la cienciometría de una forma más amplia:

> La cienciometría aplica técnicas bibliométricas a la ciencia. El término ciencia se refiere a las ciencias físicas y naturales, así como a las ciencias sociales. Pero la cienciometría va más allá de las técnicas bibliométricas pues también examina el desarrollo y las políticas científicas. Los análisis cuantitativos de la cienciometría consideran a la ciencia

como una disciplina o actividad económica. Por esta razón la cienciometría puede establecer comparaciones entre las políticas de investigación entre los países analizando sus aspectos económicos y sociales.

Los temas que interesan a la cienciometría incluyen el crecimiento cuantitativo de la ciencia, el desarrollo de las disciplinas y subdisciplinas, la relación entre ciencia y tecnología, la obsolescencia de los paradigmas científicos, la estructura de comunicación entre los científicos, la productividad y creatividad de los investigadores, las relaciones entre el desarrollo científico y el crecimiento económico, etc.

La cienciometría usa técnicas matemáticas y el análisis estadístico para investigar las características de la investigación científica. Puede considerarse como un instrumento de la sociología de la ciencia. (p. 49)

Es pues, la cienciometría, mucho más amplia que la medición de la producción en ciencia, pues con ello consigue la información necesaria que aporta a la toma de decisiones en la industria, la política, el poder, la sociedad, y más.

3.2.1.2.1. *Término correcto en español*

Manuel Krauskopf, destacado académico y científico chileno, tiene una opinión crítica con respecto al uso del término "cienciometría" en el español, como traducción directa del inglés *scientometrics*. Propuso en 1994 en revista hasta hoy líder en el tema, *Scientometrics* (Springer), el término "epistemometría", señalando una correcta etimología, que no poseería la traducción cienciometría. Indica además que el término aporta mayor claridad evitando confusiones. (Krauskopff, 1994)

Su propuesta no ha tenido el impacto suficiente para expandir el uso del término en los demás autores que publican en español. Cienciometría, cuente o no con las credenciales correctas del español, es hoy el término más ampliamente utilizado en nuestro idioma como traducción del inglés *scientometrics*.

3.2.1.3. *¿Qué es la Informetría?*

La informetría o infometría, no se escapa de estas definiciones en conjunto. Spinak (1996) la define de la siguiente forma:

> La Informetría se basa en las investigaciones de la bibliometría y la cienciometría, y comprende asuntos tales como el desarrollo de modelos teóricos y las medidas de información, para hallar regularidades en los datos asociados con la producción y el uso de la información registrada. La Informetría trata de la medición de todos los aspectos de la información, el almacenamiento y su recuperación, por lo que incluye la teoría matemática y la modelización. En sentido más amplio estudia los aspectos cuantitativos de la información, no solamente la registrada como los registros bibliográficos, sino todos los aspectos de la comunicación formal o informal, oral o escrita. (pp. 131, 132)

3.2.2. Indicadores bibliométricos o índices bibliométricos

Los indicadores o índices bibliométricos, también llamados a veces cienciométricos, tal como se mencionó en el punto 3.2.1 son fruto del entrelazamiento entre la bibliometría, cienciometría e informetría. Es por ello que aun cuando algunos autores los tratan por separado, aquí se los tratará en forma conjunta.

La cantidad de citas que recibe un artículo, habla de un aspecto a considerar en bibliometría, que puede ser relevante también en cienciometría. Así también, la cantidad de publicaciones de un autor o de una revista, habla de otro aspecto a considerar en bibliometría y cienciometría. Ambos, la cantidad de citas, y la cantidad de publicaciones, son indicadores, pues, hablan de una propiedad en particular del documento científico o de quién emana el documento —el autor o la revista—. Ambos ejemplos son indicadores mono factoriales, pues tienen solo una variable, cálculo y comprensión simple.

Existen también muchos otros con definiciones matemáticas más complejas, algunos de ellos toman en cuenta varios indicadores para en conjunto poder generar información más rica,

que entregue luces desde varios frentes, ellos aportan por ende un importante insumo a la hora de tomar decisiones más complejas.

Existen variadas críticas a la utilización de estos indicadores. En base a la discusión de diversos autores Suarez y Pérez-Anaya (2018) resumen:

> Los indicadores bibliométricos ofrecen un método estándar para la medición del desarrollo científico, aunque algunos críticos insisten en las debilidades como herramientas de evaluación al medir únicamente la producción y el impacto, pero no realmente la calidad de los procesos de investigación. No obstante, los indicadores generan información relevante sobre el proceso de investigación, volumen, evolución, visibilidad, estructura, actividad-producción, e influencia, sobre todo permiten a la institucionalidad unificar criterios para las decisiones técnicas, administrativas y políticas. (p. 97)

3.2.2.1. *Indicadores de la producción científica*

3.2.2.1.1. *Cantidad de publicaciones*

La cantidad de publicaciones, o la productividad, de un autor, una revista o institución, representa una medida cuantitativa del aporte a la comunidad científica, en cuanto se entienden las publicaciones como un aporte a la generación de conocimiento. Medir el aumento de las publicaciones en un área determinada, en bibliometría corresponde a una medida funcional u operativa de medir el crecimiento del conocimiento en tal área de estudio (Spinak, 1996, p. 80).

3.2.2.1.1. *Ley de Bradford y dispersión de las publicaciones entre las fuentes*

El químico, bibliotecario y científico de la información Samuel C. Bradford, realizó en 1934 el que sería el cuarto estudio bibliométrico de la historia, analizando la dispersión de artículos en revistas de geofísica y lubricación, enunciando lo siguiente:

> Por esto, la ley de distribución de artículos en un tema dado en revistas científicas puede establecerse de la siguiente manera: si las revistas científicas se ordenan en secuencia

> decreciente de productividad de artículos sobre un tema dado, estas pueden dividirse en un núcleo de revistas dedicadas más en particular al tema y varios grupos o zonas conteniendo el mismo número de artículos que el núcleo, donde el número de revistas en el núcleo y las zonas sucesivas estará en la relación de $1 : n : n^2$...(Bradford, 1985, p. 178)[5]

Si bien, muchos investigadores han hecho críticas con respecto a la replicabilidad de los resultados, el planteamiento original de Bradford ha perdurado, solo son algunos ajustes para su aplicación. Con respecto a la replicabilidad de los resultados Brookes (1969) concluye lo siguiente:

- El tema de la bibliografía debe estar bien definido.
- La bibliografía debe ser lo más completa posible, esto es que los artículos y revistas más relevantes deben estar incluidos.
- La bibliografía debe limitarse a un período de tiempo tal, de manera que todas las revistas que contribuyen tengan la misma oportunidad de contribuir artículos.

La distribución de las publicaciones de un tema determinado en las revistas puede verse en los bibliografos —gráficos que describen la ley de Bradford (Spinak, 1996, p. 33)—. La gran cantidad concentración de las publicaciones en unas pocas fuentes, la gran cantidad de fuentes con muy pocas publicaciones del tema, son comportamientos sociales que se pueden describir y analizar desde diferentes áreas del conocimiento como la estadística o la economía. Brooks (1969) desde la bibliometría tiene la siguiente interpretación:

> Cuando se escriben los primeros artículos de un tema nuevo se envían a una selección pequeña y adecuada de revistas y son aceptados. Estas revistas seleccionadas inicialmente atraen más y más artículos a medida que el tema se desarrolla, pero, al

[5] Trabajo original publicado en 1934, en *Engineering: An Illustrated Weekly Journal*, 137, 85-86.

> mismo tiempo, otras revistas comienzan a publicar sus primeros artículos en el tema. Si el tema continúa creciendo entonces surge un núcleo Bradford de revistas las que son las más productivas en el tema. A medida que esto ocurre, la presión del nuevo tema sobre la revista se incrementa hasta que se imponen restricciones de espacio o por decisiones editoriales de balancear con otros temas. (p. 954)

Esa dinámica de crecimiento se puede ver en los bibliografos al comparar la curva en el núcleo de Bradford con el resto de las zonas.

3.2.2.1.2. *Evolución cronológica de la producción*

Además de los análisis de productividad según cantidad y dispersión, la productividad puede medirse a través de los años, evaluando la dinámica de las revistas del tema, pudiendo proyectar su performance a futuro.

3.2.2.2. *Índices de impacto.*

Corresponden a una de las diferentes clases de indicadores bibliométricos. Su objetivo es expresar medidas que indiquen el impacto de publicaciones, autores, revistas u otras instituciones en la comunidad científica.

Importante es recalcar que, no basta con utilizar una sola métrica para cubrir todas las características consideradas relevantes en la medición de la calidad de las producciones científicas, por ello, un modelo que permita la combinación de varias de ellas de forma complementaria corresponde a una práctica adecuada para una evaluación más completa (Suarez Colorado & Pérez-Anaya, 2018, p. 113).

En los puntos siguientes se exponen los indicadores de impacto básicos utilizados en esta investigación.

3.2.2.2.1. *Índice H.*

El profesor Jorge Hirsch propone el 2005 el índice de impacto más usado hoy para medir el impacto de científicos.

Sus palabras fueron: "*I propose the index h, defined as the number of papers with citation number ≥ h, as a useful index to characterize the scientific output of a researcher.*" (Hirsch, 2005, p. 16569). Dicho de otra forma: la cantidad H de publicaciones con por lo menos una cantidad H de citas.

Para calcularlo, se ordenan en orden decreciente las publicaciones por el número de citas recibidas, enumerándolas, para identificar el punto en el que el número de orden coincide con el número de citas recibidas por una publicación, dicho número constituye el índice h. Por ejemplo: Un **índice H = 9** significa que al menos **9 artículos** han recibido **9 citas** cada uno.

Un año después comenzó a aplicarse también para medir el impacto de revistas (Braun et al., 2006) y posteriormente a otros niveles jerárquicos en la producción científica, como instituciones o países.

Este indicador no ha estado exento de críticas, sin embargo, además de no haber cesado su uso como instrumento para la medición del impacto de autores e instituciones, se han generado decenas de factores de impacto como derivaciones del índice H.

3.2.2.2.2. *Índice G.*

Propuesto por el profesor Leo Egghe (2006a) como mejora al índice H, buscando dar mayor importancia en el cálculo de la medida también al TC de los artículos más citados. La definición fue la siguiente:

A set of papers has a g-index g if g is the highest rank such that the top g papers have, together, at least g^2 *citations. This also means that the top g + 1 papers have less than* $(g + 1)^2$ *papers.* (Egghe, 2006b, p. 132)

En otras palabras, el índice G corresponde al conjunto **g** de artículos que tiene un TC de por lo menos g^2.

3.2.2.2.3. *Cociente m.*

Fue introducido por el mismo Hirsch en su clásico del 2005, solo de forma secundaria al índice H, refiriéndose a él como parámetro *m*: "*Quite generally, the slope of h versus n, the parameter m, should provide a useful yardstick to compare scientists of different seniority".(p. 16570)*

Posteriormente otros autores lo incluyeron en sus análisis simplificando su definición como $^{h}/_{y}$, siendo "h" el índice H e "y" la cantidad de años del autor publicando (desde la primera publicación), nombrándolo como **cociente *m*** *(Bornmann et al., 2008; Bornmann et al., 2011; Díaz et al., 2016).*

3.2.2.2.3.1. *Nombre equivocado en Biblioshiny.*

Un detalle importante a mencionar es que Biblioshiny[6], nombra a este indicador con la etiqueta de "*M-index"* en su plataforma web —menú *Sources/Source Impact*, *Impact measure/M-index—*, en las tablas *Source Impact* con los resultados (tablas como la), aparece como "m-index".

Más allá de las mayúsculas, el nombre está erróneo, pues el *m-index* corresponde a otro indicador propuesto el 2008 (Bornmann et al.), de hecho, fue en la misma publicación donde se menciona el cociente *m*, pudiéndose ver claramente las diferencias, pues el índice M (o *m-index*) tiene una definición completamente diferente.

De todas formas, aunque el nombre en Biblioshiny no es el correcto, los valores entregados para este indicador sí se corresponden con la definición hecha por los autores ya mencionada en el punto anterior (**3.2.2.2.3**), por lo que en este trabajo solo se procuró de modificar el nombre.

3.2.2.2.4. *Total de citas (TC).*

El conteo total de citas (TC) corresponde a una medida de impacto por sí misma. Pero el hecho de mida un solo parámetro, la hace un indicador pobre. Suele estar acompañada de otros indicadores para poder obtener análisis más completos.

3.2.3. Bases de datos bibliométricas (BDB).

Para los estudios bibliométricos, las revisiones sistemáticas, y cada vez que se requiera buscar información de fuentes acordes a una investigación científica, las bases de datos bibliométricas son fundamentales. Hoy, las tres principales bases de datos bibliométricas (BDB) son Google Scholar, Web of Science y Scopus.

[6] Aplicación web para el análisis de datos bibliométricos. Detalles sobre el software en punto 6.5.3, sobre su uso en los puntos 7.3 y 7.4.

Estos buscadores son herramientas mucho más especializadas para el estudiante, profesor o investigador que requiera información académica, ya sea de revistas científicas, capítulos de libros, y diversos tipos de documentos. Salvo casos excepcionales, la información provista por estos servicios cuenta con la identificación correspondiente de autores, años, fuentes, varios campos más y hasta la cantidad de citas y referencias, con la posibilidad de hacer seguimiento a las mismas, pudiendo navegar por las rutas de conocimiento que se tejan antes de la publicación, y cómo prosiguen estos caminos después de publicado el documento.

Investigaciones como la de Harzing y Alakangas (2016) se han dedicado a estudiar las diferentes coberturas de las diferentes BDB, cómo determinadas áreas del conocimiento se encuentran más o menos representadas en cada una. Cuál sea la BDB usada para la obtención de los datos de estudio determina finalmente los resultados de la investigación (Wanyama et al., 2022), por ello es fundamental especificar cuál es la BDB usada, o cuáles, si son más de una.

4. Planteamiento del Problema

- **Pregunta de Investigación**

¿Qué bases de datos vinculadas a la computación científica son posibles de identificar en la literatura extraída de la base de datos de *Web of Science* ―de los últimos diez años― que permitan trabajar soluciones a problemas auténticos (reales) en entornos de aprendizaje para la educación STEM?

- ¿Cuáles son los tópicos de investigación ―subdisciplinas― con mayor desarrollo en el campo de las bases de datos científicas?
- ¿Cuáles son los países y que tipo de financiamiento utilizan los investigadores para contribuir al campo de las bases de datos científicas?
- ¿Cuáles son las revistas y bases de datos que han tenido un mayor impacto en el desarrollo de la ciencia en la última década?
- ¿Qué oportunidades y desafíos es posible identificar en el uso de bases de datos científicas para la educación STEM?

5. Objetivos

5.1. Objetivo general

Caracterizar el conocimiento bibliométrico de la literatura extraída de la base de datos de *Web of Science* sobre bases de datos web vinculadas a la computación científica —de los últimos diez años— evaluando el impacto de las publicaciones en revistas con o sin trayectoria en el tema.

5.2. Objetivos específicos

1) Generar un *dataset* con registros bibliográficos de artículos científicos sobre bases de datos web, que cuente con parámetros mínimos de impacto, apto tanto para análisis bibliométrico como también para revisión sistemática.

   a) Comparar los métodos de deduplicación por EndNote X9 y por Microsoft Excel usando las referencias bibliográficas extraídas desde *Web of Science*, seleccionando el que entregue mejores resultados.

   b) Realizar un barrido por impacto que descarte los registros de artículos científicos sin impacto relevante en la temática foco de esta investigación, estén o no publicados en revistas con trayectoria en el tema.

2) Realizar un análisis bibliométrico sobre los resultados del punto uno, enfocado en evaluar el impacto de las publicaciones en revistas con o sin trayectoria en el tema.

   a) Caracterizar de forma general la literatura científica sobre las bases de datos web.

   b) Identificar las revistas científicas más influyentes en el área de las bases de datos web, midiendo relevancia, fuentes de referencia, e impacto según índice H y otros factores que permitan un análisis epistemométrico[7] más completo.

[7] Voz más apropiada para referirse al significado y las potenciales metodologías del anglicismo *scientometrics* en países de habla hispana Krauskopff, M. (1994). Epistemometria, a term contributing to express the meaning and potential methodologies of scientometrics in Spanish speaking countries. *Scientometrics, 30*(2), 425-428. https://doi.org/10.1007/BF02018117 .

6. Metodología

Para responder a la pregunta de investigación se plantea iniciar una investigación bajo los marcos metodológicos de revisión sistemática que plantea PRISMA, considerando un análisis bibliométrico sobre la temática ya detallada anteriormente. En este respecto, se procedió a realizar una búsqueda sistemática que permitió identificar y descartar duplicados. Posteriormente, se realizó un barrido inicial por impacto, el cual redujo el tamaño de la muestra generando el *dataset* final. Sobre éste *dataset*, se realizó un análisis bibliométrico que permitió caracterizar el área de estudio, considerando una evaluación para elegibilidad de las publicaciones. El trabajo derivado de esta tesis servirá como un importante insumo para proseguir con el proceso de revisión sistemática.

6.1. Búsqueda sistemática

En función de los objetivos de investigación, la búsqueda de artículos científicos que proveen bases de datos web se realizó utilizando los catálogos *Science Citation Index Expanded* (SCI-EXPANDED), *Social Sciences Citation Index* (SSCI) y *Arts & Humanities Citation Index* (A&HCI) que constituyen la colección principal de la base de datos bibliográfica *Web of Science* (WoS). En el proceso de búsqueda se emplearon las palabras claves: *database** en el campo título y *web** en tema, ambos de forma excluyente. El periodo de tiempo considerado fue la última década, desde el 01 de enero de 2012 al 16 de mayo de 2022. En la Figura 1, se muestran los campos, parámetros y operadores booleanos empleados en la búsqueda:

La Tabla 2 muestra los criterios empleados en la búsqueda, los cuales pueden volver a

database* (Title) and **web*** (Topic) and **Articles** or **Review Articles** or **Data Papers** or **Database Reviews** (Document Types) and **English** or **Spanish** (Languages) and **Science Citation Index Expanded (SCI-EXPANDED)** or **Social Sciences Citation Index (SSCI)** or **Arts & Humanities Citation Index (A&HCI)** (Web of Science Index) and **Timespan: 2012-01-01 to 2022-05-16 (Index Date)**

Figura 1. Parámetros de búsqueda utilizados en Web of Science.

consultarse en el siguiente enlace:

https://www.webofscience.com/wos/woscc/summary/89ce8fef-702b-4712-b0c8-a0ae993de4ca-386929d4/times-cited-descending/1.

Tabla 2. *Búsqueda en WoS.*

Colección e índices	Tipo de documento	Idiomas	Intervalo de tiempo (según fecha de indexación)	Términos de búsqueda(1)	Resultados	Fecha de la búsqueda
WoS-CC: SCI-EXPANDED, SSCI, A&HCI	Artículos, *reviews*, *data papers* o *reviews* de base de datos	Inglés y español	2012-2022(2)	*database** (Título) y *web** (Tema)	2565	17-05-2022

(1): Los asteriscos (*) al final de los términos de búsqueda, no hacen referencia a alguna nota en pie, únicamente para las Tabla 2, Tabla **3** Tabla 4 y Tabla 5, los asteriscos corresponden a un caracter propio de la estrategia de búsqueda (Clarivate Analytics, 2022)

(2): La fecha precisa de corte es el 16 de mayo del 2022

Para la descarga de los registros se consideraron los 29 campos disponibles en WoS (Figura 1). El proceso de descarga se llevó a cabo en tres pasos —1000 registros en cada paso—, considerando los siguientes formatos: i) EndNote Desktop (.ciw); ii) BibTex (.bib), iii) Excel (.xls) y iv) texto (.txt). Posteriormente, los registros fueron compilados en cada tipo de formado en un solo archivo.

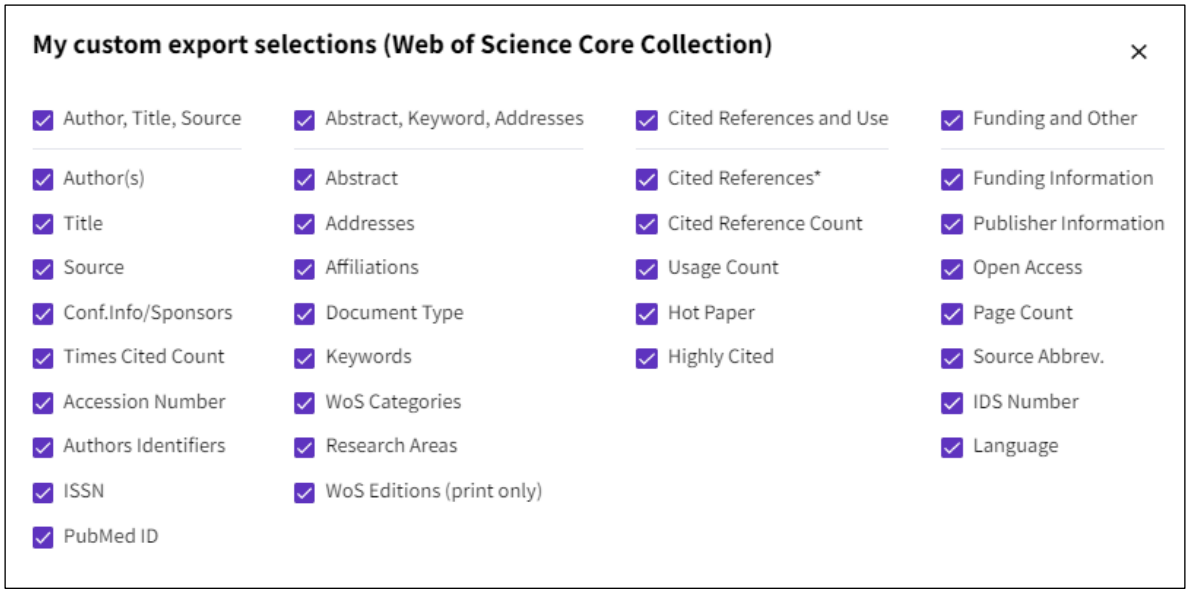

Figura 1. Campos seleccionador al momento de guardar los resultados de búsqueda

6.2. Primer barrido: Deduplicación

6.2.1. Detección de duplicados usando EndNote X9.

Los resultados de búsqueda obtenidos se sometieron a un proceso de detección y descarte de duplicados —deduplicación—. Primero con el gestor de referencias EndNote X9, con la

opción de búsqueda de duplicados, en menú *References*/*Find Duplicates,* con las preferencias predeterminadas para la detección de duplicados: *Author, Year, Title y Reference Type*.

6.2.2. Detección de duplicados usando Excel.

El proceso de detección de duplicados en Excel, consideró la búsqueda de coincidencias del número DOI usando la función de formato condicional por columna según la siguiente serie de pasos:

1) Se seleccionó la columna DOI, aplicando en menú *Inicio/Estilos/Formato condicional/Reglas para resaltar celdas/Valores duplicado.*
   Se consideraron duplicados los registros con igual DOI, por lo que se debió completar los DOI de los registros en blanco.
2) Para los registros cuyo DOI no fue encontrado, se procedió a filtrar por título usando formato condicional en esta columna, cautelando la información de registro de ambos campos.

6.2.3. Base de datos con registros únicos.

La construcción de una base de datos con registros únicos se realizó cotejando la cantidad de citas de cada registro duplicado, para esto se mantuvo el registro con mayor cantidad de citas y se incorporaron las citas del registro descartado, si corresponde.

6.3. Generación de un nuevo índice F_H.

Se definió el índice de impacto "$índice\ F_H$" a partir del índice H y del total de citas (TC) anualizados, para lo cual se consideró:

$$Total\ de\ citas\ anualizado\ (TCA) = \frac{TC}{\Delta t + 1} \qquad \textbf{(1)}$$

siendo $\Delta t + 1$ el transcurso entre que la revista haya comenzado a publicar en el tema y el año actual, tomando en cuenta ambos años;

$$índice\ F_H = índice\ H \cdot TCA \qquad \textbf{(2)}$$

Para el cálculo del $índice\ F_H$ se empleó la base de datos con registros únicos, la cual fue utilizada para obtener los índices H, cociente *m*, TC, cantidad de publicaciones (N) y año de inicio de publicaciones, usando la aplicación *Biblioshiny*.

6.4. Análisis bibliométrico.

6.4.1. Selección de registros considerando índice F_H.

Para el análisis bibliométrico se seleccionaron los registros que tuviesen $índice\ F_H \geq 2$ para cada revista. Posteriormente, se realizó un análisis usando Biblioshiny, que incluyó:

a) Descripción general y comparativo del *dataset* desde la tabla *Main Information*, evaluación del efecto de aplicar el índice F_H para la selección por impacto

b) Análisis del *dataset* con foco en el nivel[8] de fuentes (revistas), midiendo relevancia según producción, impacto por citas e impacto según índice H y derivados.

6.5. Software

6.5.1. EndNote.X9.

Como gestor de referencias se usó EndNote X9 (The EndNote Team, 2013), en su versión de escritorio, tanto para las referencias del propio documento de tesis como en el proceso de deduplicación del *dataset* (puntos 6.2 y 7.2.1)

6.5.2. Excel.

Las funciones de filtros y formato condicional mencionadas en el proceso de deduplicación del *dataset* (puntos 6.2 y 7.2.2) fueron probadas en las versiones de Office 2016, 2019 y 365 (Microsoft Corporation, 2018).

[8] Bibliometrix puede realizar análisis en niveles de fuentes, autores y documentos. https://bibliometrix.org/biblioshiny/assets/player/KeynoteDHTMLPlayer.html#37.

6.5.3. Bibliometrix.

El análisis bibliométrico —posterior al proceso de deduplicación— fue hecho mediante la plataforma web de acceso abierto Biblioshiny, interfaz gráfica para el paquete R *bibliometrix (Aria & Cuccurullo, 2017)*.

Se trabajó con la versión de R y RStudio 4.2.1 (RStudio Team, 2020), allí se instaló el paquete *bibliometrix (*v.3.2.1, más tarde su actualización v.4.0.0, esto no implica mayores diferencias en los cálculos realizados), y se lanzó la aplicación web Biblioshiny.

Una vez cargado el *dataset* a analizar, la plataforma Biblioshiny realiza los cálculos en sus propios servidores, por lo que, las características del ordenador (RAM o procesador) no son de gran relevancia.

7. Resultados y Discusión

7.1. Resultados de la búsqueda

7.1.1. Diseño de búsqueda

7.1.1.1. *Primeras aproximaciones y variación de la cantidad de resultados obtenidos en la búsqueda sistemática "N".*

Con el objetivo de apuntar al área de las "bases de datos vinculadas a la computación científica" planteada en el objetivo general de este trabajo, se optó por usar en la búsqueda sistemática el término "*database"* —de forma excluyente— en el título de las publicaciones. Para que los artículos científicos resultantes de esta búsqueda hicieran referencia a la computación científica como área, se optó por incluir términos asociados a esta —ej. bioinformática—. El conjunto completo de parámetros de búsqueda y el total de resultados para este primer término sumado a *database* se muestra en la Figura 2.

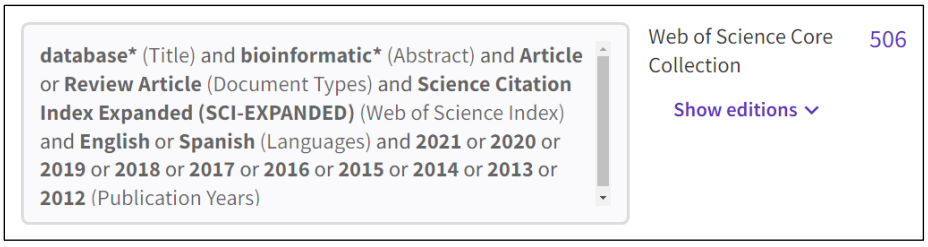

Figura 2. Parámetros de la primera búsqueda. Esta incluye los términos de búsqueda *database** en título y *bioinformatic** en *abstract*. Captura tomada desde el historial de búsqueda del investigador (https://www.webofscience.com/wos/history)

Luego, para incluir otros campos de la ciencia que también entraran bajo el concepto de computación científica, se añadieron los conceptos de biología computacional, quimio informática y química computacional. Esta tarea fue realizada en búsquedas independientes bajo los mismos parámetros de búsqueda ya mostrados en la Figura 2, cambiando solo el término usado en el campo[9] de búsqueda "Abstract" (ver Tabla 3).

[9] No confundir con campo de la ciencia.

Tabla 3. Número de registros para las primeras cuatro búsquedas independientes.

Término de búsqueda en campo		N
"Title"	***"Abstract"***	
*database**	*bioinformatic**	506
*database**	*cheminformatic**	31
*database**	*computational biology**	49
*database**	*computational chemistry**	30
Total (incluye posibles duplicados)		**616**

Los asteriscos al final de los términos de búsqueda fueron incluidos para considerar conceptos derivados que pudiesen terminar con otra serie de caracteres (Clarivate Analytics, 2022). La Tabla 4, muestra los resultados obtenidos al realizar la búsqueda incluyendo el asterisco al final del primer término en el caso de los conceptos en bigrama[10], pasando de incluir solo resultados con el término *computational* (computacional), a variaciones en su terminación como *computationally* (computacionalmente).

Tabla 4. Número de registros al incluir asterisco en el primer término de bigramas.

Término de búsqueda en campo		N
"Title"	***"Abstract"***	
*database**	*bioinformatic**	506
*database**	*cheminformatic**	31
*database**	*computational* biology**	54
*database**	*computational* chemistry**	33
Total (incluye posibles duplicados)		**624**

Para ampliar aún más la búsqueda a otros sub campos de la ciencia que no lleven necesariamente en su título los conceptos de "informática" o "computacional", pero que hagan uso de bases de datos disponibles en la web y por tanto estar desarrollando "computación científica", se optó por incluir otros siete sub campos de la ciencia.

Además, se buscó no solo en el campo *abstract*, también en *topic*, para incluir en los resultados publicaciones que incluyesen estos conceptos en otros campos del documento. Dichos resultados se muestran en la Tabla 5, los parámetros de búsqueda para la búsqueda en la

[10] Concepto en dos palabras.

primera fila de la tabla se muestran en la Figura 3, los mismos fueron usados para el resto, cambiando solo el término en *topic*.

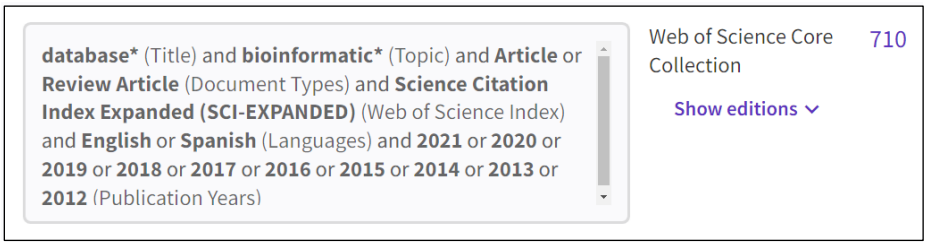

Figura 3. Búsqueda de database* en título y bioinformatic* en topic. Captura tomada desde el historial de búsqueda del investigador (https://www.webofscience.com/wos/history)

Tabla 5. Cantidad de resultados cambiando *abstract* por *topic* y añadiendo más términos.

Término en *"Title"*	**Términos en *"Topic"***	**N**
*database**	*bioinformatic**	710
*database**	*cheminformatic**	47
*database**	*computational biology*	29
*database**	*computational chemistry*	12
*database**	*drug design*	93
*database**	*drug discovery*	259
*database**	*ecosystem*	277
*database**	*environmental**	770
*database**	*material science*	26
*database**	*nanomaterial**	23
*database**	*nanotechnology*	10
Total (incluye posibles duplicados)		**2256**

¿Qué otros subcampos de la ciencia pudiesen hacer uso de las bases de datos y publicar por ende artículos sobre "bases de datos vinculadas a la computación científica? Esta pregunta motivó a incluir más subcampos de la ciencia en los términos buscados en el campo *topic*. Esto sumado a la disyuntiva de cuándo incluir asterisco en el concepto, si esto desvirtuaba la búsqueda o no, implicó un aumento progresivo del N (la cantidad de resultados de búsqueda) cada vez obtenido, sin contar la variación implicada por la inclusión o no del asterisco en cada término.

7.1.1.2. *Sesgo no deseado.*

- ¿Cómo identificar los subcampos de la ciencia más propensos a hacer uso de bases de datos disponibles en la web?

Al intentar identificar algunos de los subcampos de la ciencia que pueden asociarse al uso de bases de datos, se incluye en la búsqueda un sesgo no previsto. Si bien, al ir agregando cada vez más términos en el campo *topic* para incluir más áreas de la ciencia e ir reduciendo este sesgo cada vez más, intentar reducirlo al mínimo, requeriría agregar todos los sub campos de la ciencia, pues es imposible definir de forma clara e indiscutible cuáles campos pueden o no vincularse a las bases de datos.

7.1.1.3. *Cero sub campos, todos los campos y computación científica.*

Para cumplir el objetivo del diseño de una búsqueda cuyos resultados respondan a artículos científicos sobre "bases de datos vinculadas a la computación científica", se decide por no apuntar a ningún campo o subcampo de la ciencia, para no ir sesgando desde un inicio la búsqueda y sus resultados, pudiendo apuntar a priori, a todas las áreas de la ciencia.

Lo anterior resuelve el problema del sesgo, pero reabre uno anterior: ¿Cómo enfocar la búsqueda sobre las bases de datos a una "vinculación con la computación científica"

Si entendemos por "computación científica", como el uso de computadores para resolver problemas científicos, y que esta investigación tiene como temática principal las "bases de datos" —herramientas propias de la computación—, cada artículo científico, que incluya en su título el término base de datos, muestra ya una vinculación entre las "bases de datos" y la "computación científica".

Al realizar la búsqueda, únicamente con el concepto "bases de datos", los resultados obtenidos son más de 25.000, pero más allá del altísimo número, la especificidad sigue siendo muy baja, pues, aunque la idea ya queda clara, y es apuntar a todas las áreas de la ciencia, hay un detalle clave en el contexto que da origen a esta investigación, y es que estas bases de datos vinculadas a la computación científica, puedan ser usadas eventualmente en el diseño de actividades de aprendizaje. Algo que puede ayudar significativamente en esta tarea, es que estas bases de datos se encuentren disponibles en la web. Para cumplir este requerimiento recién mencionado, es que se decide por incluir en el campo *topic*, el término web*.

7.1.1.4. Ampliación del intervalo de búsqueda en *Index Date.*

Con el sentido dar un carácter más actual a los resultados, se decide incluir artículos científicos del año en curso. Para esto, se amplía el intervalo de búsqueda hasta ese mismo momento, fijando como nueva fecha de corte el 16 de mayo del 2022.

Al cambiar el límite superior del intervalo de búsqueda, con el objetivo de incluir publicaciones indexadas en WoS que aún no cuenten con fecha oficial de publicación —por ser de acceso temprano—, se cambia el parámetro *Publication type* por *Index type*.

7.1.2. Búsqueda final

En base a los parámetros de búsqueda ya explicitados en la sección Metodología —punto 7.1—, los resultados de búsqueda fueron 2565 registros bibliográficos de publicaciones, correspondientes a diferentes tipos de documentos científicos, siendo mayoritariamente artículos científicos. Un mayor análisis bibliométrico se muestra en el punto 8.4.2, Tabla 14, columna “Resultados - Pre barrido”.

Estos resultados, además de incluir registros en duplicado, mostraron un gran número de publicaciones con poco o nulo impacto. En base a estos dos aspectos, se ideo un proceso de depuración del *dataset* original, que incluyó dos barridos, uno por deduplicación (punto 8.2) y otro por impacto (punto 8.3).

El diseño y ejecución de la búsqueda sistemática, más la remoción de registros por deduplicación e impacto, constituyen la primera parte de una revisión sistemática. Desde ese tenor, la muestra un diagrama PRISMA con el flujo del proceso realizado (Page et al., 2021).

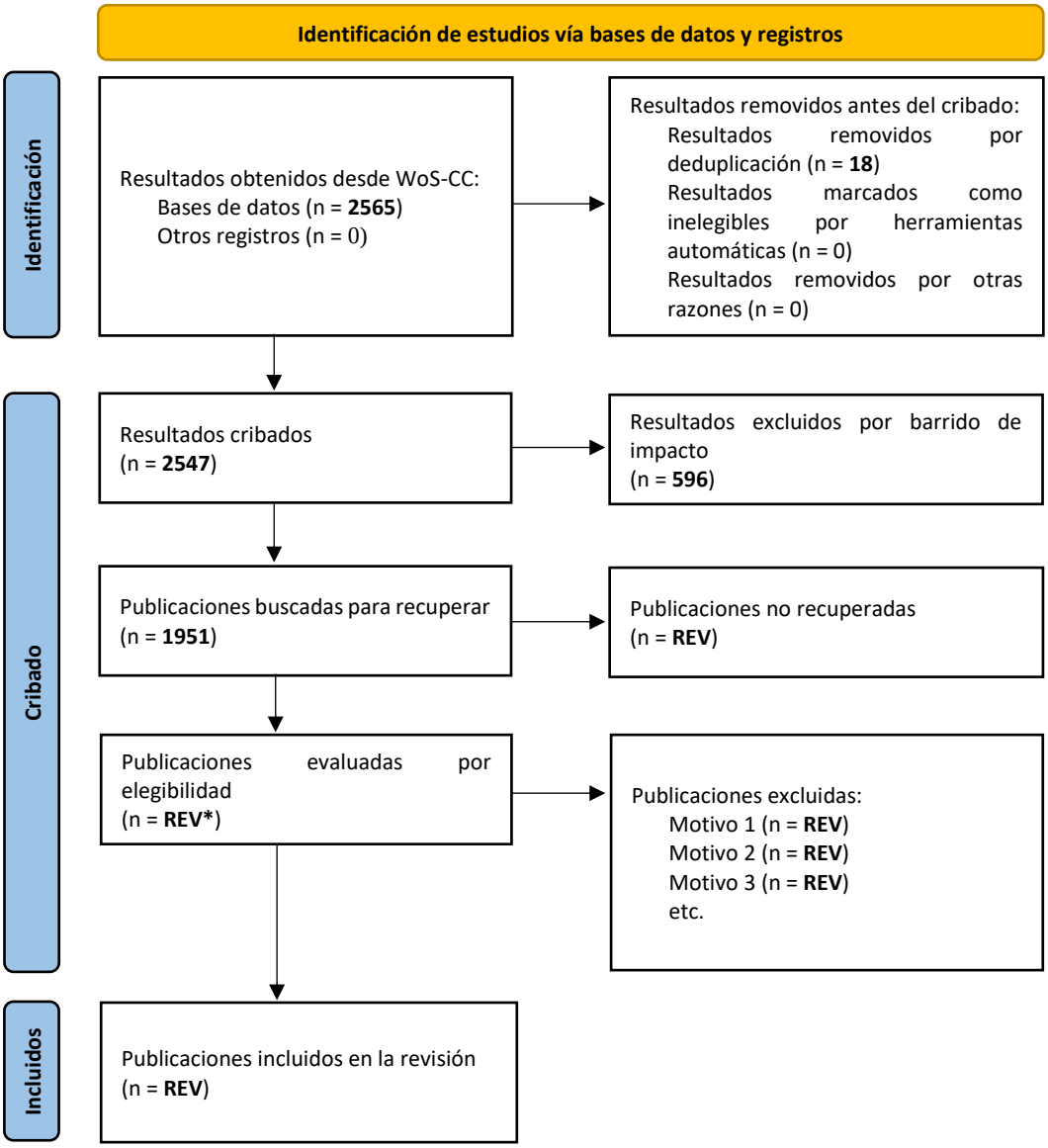

Figura 4. Diagrama de flujo según la guía PRISMA 2020

Los valores de 'n = REV' deberán ser completados a posterior con la continuación de la revisión sistemática (posterior a esta tesis).

7.2. Primer barrido: Deduplicación

7.2.1. Detección de duplicados por EndNote X9.

Al haber realizado una búsqueda en un solo paso en el buscador de WoS, los resultados no deberían contener duplicados, lo que suele pasar cuando se juntan bases de datos hechas con diferentes parámetros de búsqueda —sobre todo con términos de búsqueda relacionados entre

sí—, o en bases de datos distintas, como WoS y Scopus, pero este no es el caso. Sin embargo, al cargar los resultados en el gestor de referencias EndNote, y aplicar la opción de búsqueda de duplicados (menú *References*/ opción *Find Duplicates*), sí arrojó un total de 32 registros en duplicado (16 pares).

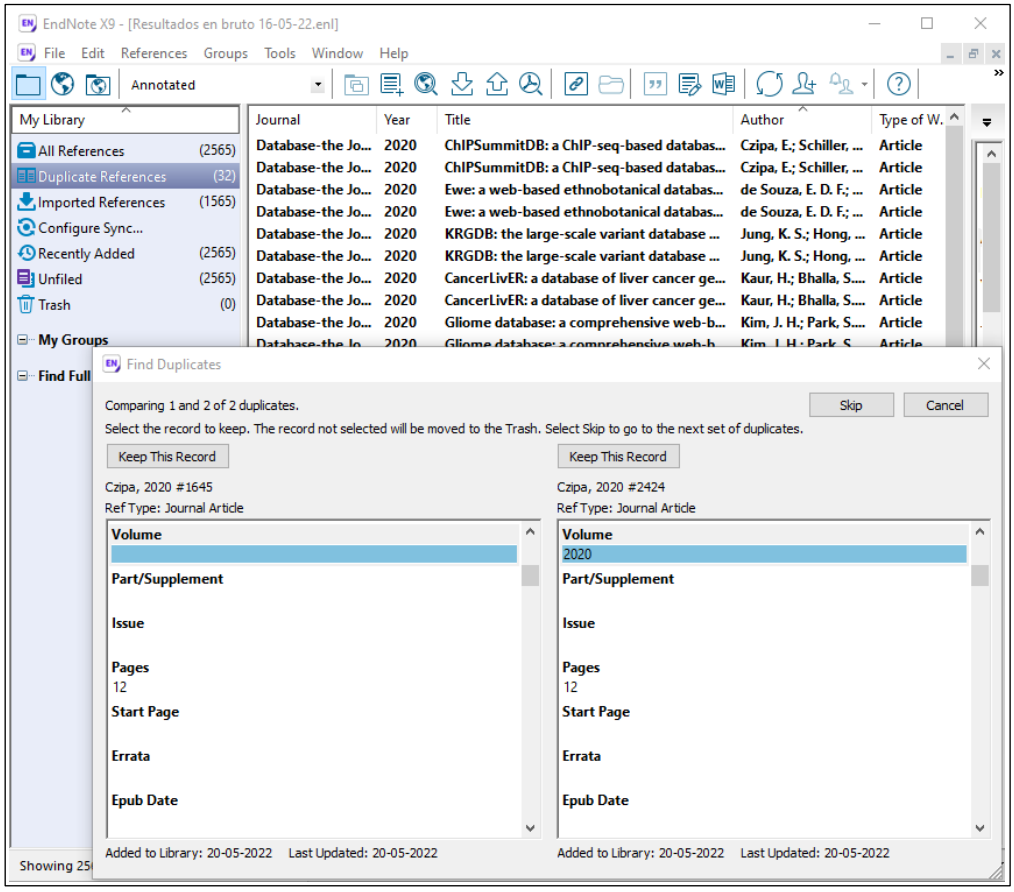

Figura 5 - Vista del gestor de referencias EndNote X9, menú References/Find Duplicates. Se muestra la cuenta de los 32 duplicados encontrados por este método y una ventana que resalta el o los campos que sí difieren si es que los hay.

EndNote separa estos registros del resto, las muestra en una lista, y en una ventana más pequeña señala las diferencias entre cada referencia y su par duplicada, para que el investigador discierna cuál guardar y cuál descartar. Las diferencias consistieron en los campos *keyword*, *abstract*, *author Address* (dirección del autor), *notes* (notas), *volume*, *accession number* (código interno de WoS) y *date* (fecha de publicación). En los dos primeros campos mencionados las diferencias se consideraron menores y poco concluyentes: en *keywords*, la diferencia fue por cantidad, uno o dos más o menos, mientras que en el resumen las diferencias consistieron sólo en errores ortográficos o de redacción. Es en los cuatro últimos campos mencionados en que

donde se apreció una diferencia más marcada: en volumen, una referencia tenía el campo en blanco, mientras que su par tenía el valor 2020; el código interno para WoS era de tipo WOS:**0005**XXXXXXXXXXX para las que tenían vacío el campo volumen, y de tipo WOS:**00077669**XXXXXXX para el segundo grupo, la mayoría de las citas estaban indexadas en los registros del primer grupo, y no en el segundo. Por último, todos los registros en duplicado según EndNote, corresponden a la misma revista: *Database: The Journal of Biological Databases and Curation*, y corresponden a publicaciones del año 2020. Para analizar con más claridad estos campos se revisaron los *datasets* completos descargados desde la búsqueda inicial, las tres partes se unieron en un solo archivo y se filtró por la revista mencionada y el año 2020, obteniendo 34 registros en duplicado al ordenar la tabla por DOI (un par más que en EndNote). Las diferencias entre ambos grupos se aprecian en la Tabla 6:

Tabla 6. Diferencias en grupos de duplicados detectados por EndNote.

	Volume	***Publication Date***	***UT (Unique WOS ID)***	**Total de citas por grupo**
Grupo 1	-	Fecha real según la publicación original	Tipo WOS:**0005**XXXXXXXXXX	28
Grupo 2	2020	"Jan 1" ó la misma que su par	Tipo WOS:**00077669**XXXXXXX	5

* Esta tabla completa se encuentra como anexo al final de este documento **(**Tabla 26**)**.

Primero, el campo *Volume*, con el valor 2020 para el grupo dos, coincide con el año de las publicaciones, por costumbre de esta revista, que edita un volumen por cada año, por ello asigna así el campo volumen, y no con el común correlativo 1, 2, 3, etcétera. Segundo, el campo *Publication Date*, coincide con la fecha real de publicación en todo el grupo uno, mientras que la mayoría de registros del grupo dos indica *Jan 1* (enero 1), de forma errónea. Lo tercero es el código interno de WoS, con un comienzo característico según cada grupo (ver en Tabla 6 el comienzo general y en la Tabla 7 los códigos completos). Estas tres diferencias recién mencionadas sirven para diferenciar ambos grupos, pero no son de gran importancia para preferir uno u otro. Por último, la cantidad de citas, siendo el grupo uno el que registra la mayor cantidad. Los detalles por campo pueden verse con distinción por pares en la Tabla 7.

Tabla 7. Diferencias en pares de registros duplicados detectados por Endnote (extracto).

Article Title	TC*	Publication Date	Publication Year	Volume	DOI	UT (Unique WOS ID)
BarleyVarDB: a database of barley genomic variation	5	NOV 28	2020	—	10.1093/database/baaa091	**WOS:0005**97231900001
BarleyVarDB: a database of barley genomic variation	0	NOV 28	2020	2020	10.1093/database/baaa091	**WOS:0007766 9**4900003
CancerLivER: a database of liver cancer gene expression resources and biomarkers	3	MAR 7	2020	—	10.1093/database/baaa012	**WOS:0005**21180200001
CancerLivER: a database of liver cancer gene expression resources and biomarkers	1	**JAN 1**	2020	2020	10.1093/database/baaa012	**WOS:0007766 9**5200020
ChIPSummitDB: a ChIP-seq-based database of human transcription factor binding sites and the topological arrangements of the proteins bound to them	4	JAN 14	2020	—	10.1093/database/baz141	**WOS:0005**21180700001
ChIPSummitDB: a ChIP-seq-based database of human transcription factor binding sites and the topological arrangements of the proteins bound to them	0	**JAN 1**	2020	2020	10.1093/database/baz141	**WOS:0007766 9**5200084
ctcRbase: the gene expression database of circulating tumor cells and microemboli	3	APR 15	2020	—	10.1093/database/baaa020	**WOS:0005**27712000001
ctcRbase: the gene expression database of circulating tumor cells and microemboli	0	**JAN 1**	2020	2020	10.1093/database/baaa020	**WOS:0007766 9**5200025

* Total de citas

** Esta tabla completa se encuentra como anexo al final de este documento (Tabla 27).

De los campos presentes en la tabla anterior, el de mayor importancia para este estudio es el de las citaciones de la publicación, por lo que fue lo principal para decidir qué registros dejar y cuáles no. Según la tabla anterior, el grupo uno, suma un total de 28 registros, mientras que el grupo dos solo cinco, por lo que se dejó los registros del grupo uno, las coincidentes a los códigos de tipo WOS:**0005**XXXXXXXXXXX. Si bien el criterio para el descarte de duplicados ya queda claro, surge un problema: las citas se presentan mayormente en el grupo uno, pero no únicamente en él, por lo que al eliminar los registros del grupo dos, podrían eliminarse también las citas allí registradas, y en tal caso, dejar dichas publicaciones con un registro de citas menor al real. Por ello, antes de eliminarlas, se procedió a revisar una por una las citas de cada publicación en cada par de registros duplicadas, para verificar si corresponden a publicaciones diferentes (publicaciones en que ha sido citado) o a citas doblemente registradas. Esta información no aparece en los metadatos descargados desde WoS, pero sí puede verificarse en él, al ingresar a la URL donde está indexada la publicación, en este caso a través del código interno de WoS —*Accession Number* en EndNote ó *UT (Unique WOS ID)* en la hoja de cálculo—.

7.2.2. Detección de duplicados por Excel.

En el punto anterior se comentó la identificación de un duplicado que no había sido identificado por Endnote, al analizar el *dataset* con Excel según las coincidencias mostradas entre los duplicados por EndNote. En este sentido, es importante preguntarse: ¿Qué tan confiable es el método de deduplicación ofrecido por un gestor de referencias bibliográficas asociado directamente a Web of Science como EndNote X9? ¿Puede ser más eficaz la deduplicación con una herramienta básica como Microsoft Excel?

El análisis de la hoja de cálculo, en Excel, tiene más de una forma de identificar valores en duplicado, uno más automatizado, diseñado específicamente para deduplicación, y el otro no, estos son descritos a continuación:

a) La primera opción consiste en ir al menú *Datos/Herramientas de datos/Detección de duplicados/Eliminar las filas con valores en duplicado*. En este punto si no se especifican los campos en donde busque duplicados, lo hará todos los campos. Esto puede realizarse de forma más certera al especificar cuáles son las columnas en donde la herramienta buscará datos en duplicado.

En este caso, al usar esta herramienta en nuestro *dataset*, al especificar un barrido por las columnas autor, título, y DOI, encuentra y elimina 16 duplicados, si busca sólo por Autor y DOI, elimina 17 duplicados, y si busca solo por DOI elimina 76 duplicados. Si bien, es el DOI el código de identificación único para una publicación científica publicada (DOI Chile, s.f.), la diferencia entre la cantidad de duplicados solo basándose en este metadato y los barridos anteriores cuando se consideró autor y título, más que triplica la cantidad inicial, sin mencionar que no se muestra cuáles fueron las filas eliminadas, por lo que termina siendo un descarte casi a ciegas, sin saber cuáles fueron los registros eliminados. Es por esto, que se revisa el segundo método para la identificación de duplicados.

b) La segunda opción, consiste en ocupar la herramienta de formato condicional en menú *Inicio/Estilos/Formato condicional/Reglas para resaltar celdas/Valores duplicados*, seleccionando primero la columna del campo DOI. Primero se aplica para identificar registros con ausencia de DOI, se completan los que sí se puedan recuperar desde las fuentes originales para luego repetir el filtro. Los registros que no posean DOI, se cotejan también aplicando formato condicional, pero en el campo título, revisando el resto de

campos en los casos que sí resalten celdas por duplicación de este campo. El flujo de este procedimiento y sus resultados puede verse en la Figura 6.

Al ordenar dicha columna por color, para ver primero las celdas destacadas, 95 de ellas son destacadas, 59 de ellas corresponden a celdas en blanco, y las otras 36 a valores del DOI duplicados. De los 59 registros sin DOI, 34 no contaban con dicho metadato[11] en sus fuentes originales, luego, al revisar duplicación aplicando formato condicional por título, no se encontraron duplicados reales, solo un grupo de registros con igual título que correspondían a la publicación de una base de datos y las posteriores actualizaciones de dicho artículo, que fueron publicadas bajo el mismo nombre[12].

De los 59 registros sin DOI, 25 sí contaban con él en sus fuentes originales, se completó este dato en el *dataset* y se repitió el filtro por formato condicional al campo DOI. 36 fueron nuevamente las resaltadas, correspondientes a 18 pares de registros en duplicado. 16 pares además de coincidir en DOI, coincidieron en título y en autores —entre otros campos—, este grupo de 16 duplicados corresponde a los ya encontrados por EndNote (punto 7.2.1). Un par corresponde a una publicación de la misma revista de la que provienen los duplicados detectados en el método por EndNote, pero no coincide en título, ya que contiene un pequeño error ortográfico en un registro y en la otro no. El último par de registros resaltados por DOI coincide también en título, pero no en autores, ya que en un registro aparece un autor menos que en el otro (ver Tabla 8).

Al comparar este método del formato condicional "b)", con el de la herramienta de detección de duplicados de Excel "a)", se desprende lo siguiente: El método "a)" detectó 76 duplicados al filtrar solo por DOI, mientras que en b) se encontraron al final solo 18, pero sumados a 58 de eliminar considerando 1 verdadero del grupo de los 59 sin DOI, se iguala el conteo de los 76.

[11] "Dato sobre otro dato". En este caso, corresponde a la información que debiese contener alguno de los campos, entregando información sobre algún parámetro de el ítem descrito en cada una de las filas (registros bibliográficos).
[12] Este grupo de publicaciones sí requirió de ciertas correcciones al contener errores en el conteo de citas. Se explica con mayor detalle en el punto 7.3.6.2.

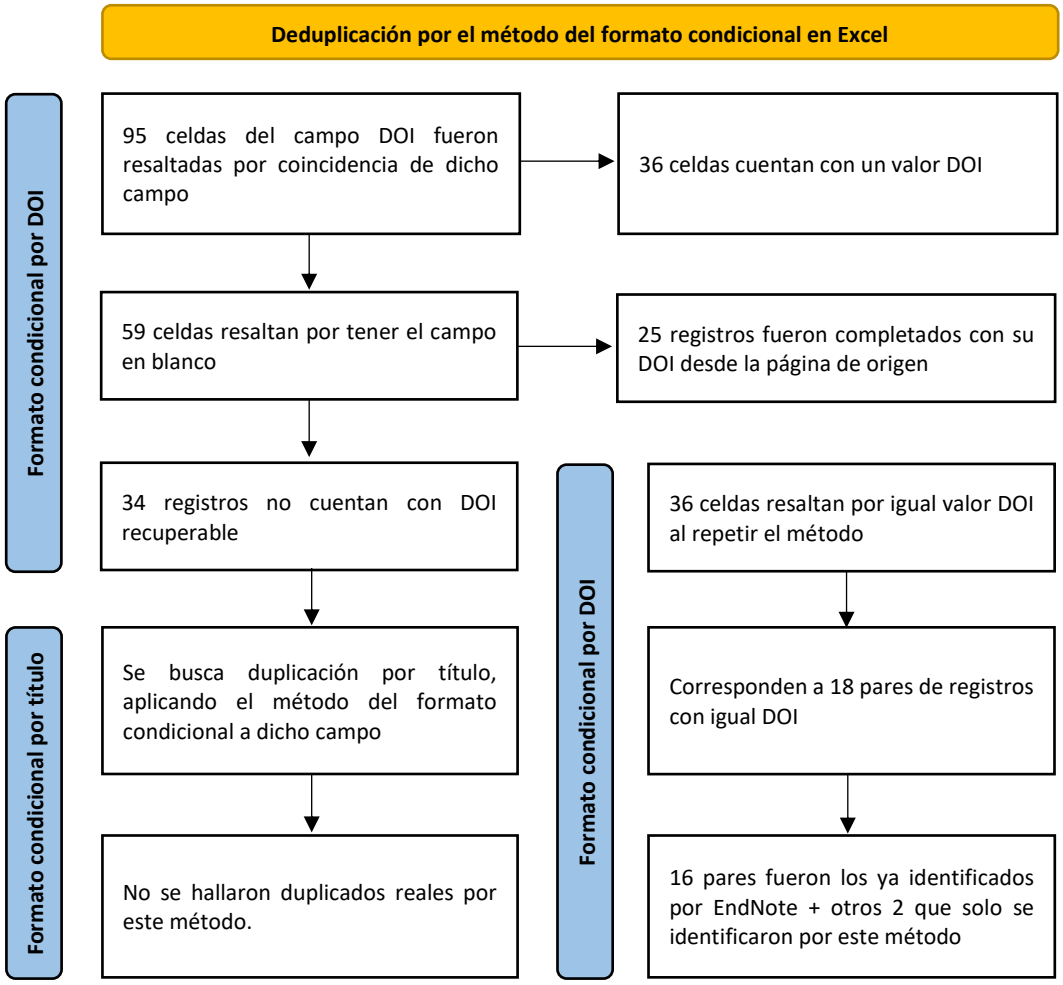

Figura 6. Diagrama de flujo para el proceso de deduplicación por Excel según formato condicional

Tabla 8. Coincidencias según campo para los duplicados encontrados por formato condicional

Pares de registros	Coincidencias por campo			Detectados por EndNote
	DOI	TI	AU	
16	Sí	Sí	Sí	Sí
1	Sí	Sí	No	No
1	Sí	No	Sí	No

Muchos de los campos que aparecen con celdas en blanco pueden ser de una importancia menor, pero en ciertos casos, como en el campo del DOI, valor de identificación único que permitió en el caso de los 18 pares de registros duplicados corroborar que se trata de la misma publicación, aun cuando el código interno de WoS dijese lo contrario, podría interpretarse como la falta de información importante.

7.2.3. Descarte racional de duplicados.

Se eliminaron las 18 filas de los valores duplicados, ajustando los valores de citas en:

a) (Kaur et al., 2020), código interno WOS:000521180200001, modificando de 3 a 4 citas, considerando la cita adjudicada al duplicado de código WOS:000776695200020.
b) (Jung et al., 2020), código interno WOS:000521177500001, modificando de 11 a 14 citas, descartando 2 citas en el duplicado de código WOS:000776695200043, por ser solo referencias desde el artículo de tipo corrección (con diferente DOI: 10.1093/database/baaa030) de código WOS:000776695200093, el que sí tenía 3 citas para sumar.
c) (Luo et al., 2020), código interno WOS:000545997600001, modificando de 0 a 1 cita, considerando la cita adjudicada al duplicado de código WOS:000776695200050.
d) (Spoor et al., 2020), código interno WOS:000548859200001, modificando de 0 a 1 cita, considerando la cita adjudicada al duplicado de código WOS:000776695200016.

Se aplicaron los cambios de DOI, deduplicación y cantidad de citas en el *dataset*, tanto en la hoja de cálculo como en los archivos de texto y formato EndNote.

7.3. Segundo barrido: Selección por índice de impacto

7.3.1. Primera aproximación: Biblioshiny, *Source Impact* e índice H.

7.3.1.1. *Carga de archivos en Biblioshiny. Carga de archivos en Biblioshiny.*

Al querer analizar el *dataset* —ya sin duplicados—, y descartar de allí las publicaciones menos relevantes, como parámetro de impacto se optó por tomar el índice H interno de la revista, aquel relativo solo al tema de esta investigación, las bases de datos web. Este índice se obtuvo desde Biblioshiny, una aplicación web para el uso de la librería *Bibliometrix (Aria & Cuccurullo, 2017)*, con scripts diseñados específicamente para el análisis bibliométrico de bases de datos de registros bibliométricos, ocupando el archivo de texto de la base de datos post deduplicación, según el siguiente método:

- Desde el *software* RStudio, se cargó la librería *bibliometrix* con

  > library(bibliometrix)
- Se lanzó la interfaz gráfica Biblioshiny con

> biblioshiny()

- En Biblioshiny se cargó la base de datos a analizar seleccionando:
  a) menú principal *Data/Load data*
  b) en el submenú *Import or Load Files/Please choose want to do*, se especificó el tipo de archivo *Import raw file(s)*
  c) en el submenú *Database/* se especificó el tipo de base de datos con que se trabajó, *Web of Science (WoS/WoK)*
  d) luego para subir el archivo, en el submenú *Choose a file/Browse*, se cargó el archivo de texto de la base de datos post deduplicación. Al ver la leyenda "*Upload complete*" se entienda el proceso de carga terminado
  e) para someter al análisis de los datos por parte de Biblioshiny clic en "*Start*".

Bajo el botón "Start", en "*Conversion results*", se indica la cantidad de registros reconocidos en el archivo cargado, este número debe coincidir con la cantidad que se haya trabajado previamente. A la derecha, se mostrará toda la información contenida en el archivo como una tabla de datos. (Ver Figura *7*)

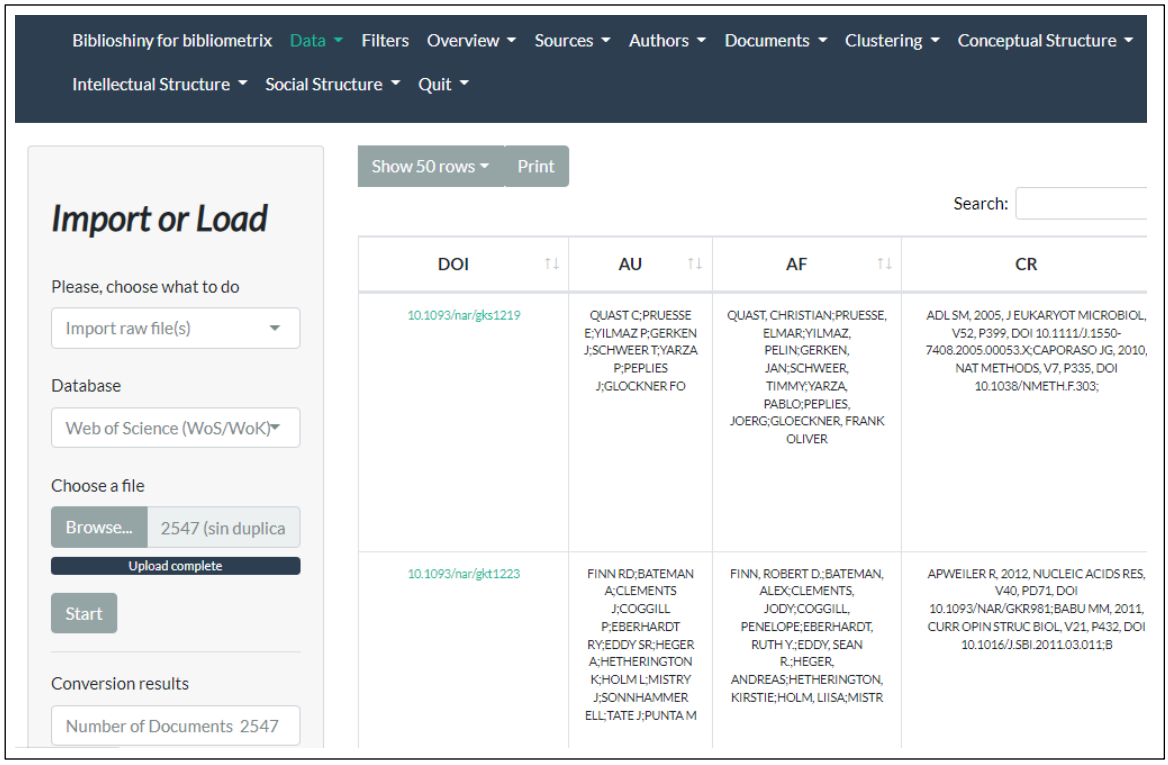

Figura 7. *Dataset* en Biblioshiny.

7.3.1.2. *Obtención de la tabla Source Impact e índice H.*

El índice H interno de la revista, se obtuvo desde el menú *Sources/Source Impact, Impact measure/H Index*, clic en "Start". Esta orden entrega un gráfico del impacto según el índice H interno de la revista, pero solo para un máximo de 100 revistas, por lo que para analizar con más detalle la información obtenida, arriba del gráfico se seleccionó *Table/Show All/Excel*, para descargar la tabla completa en formato hoja de cálculo.

La tabla *Source Impact* descargada (tabla de impacto de las revistas) muestra un total de 856 revistas, en ella pueden verse tres índices para la revista, el índice H y dos de sus derivaciones, el índice G y el cociente *m*. Además, tres datos complementarios referentes también a cada revista: el total de citas acumulado (TC), la cantidad de publicaciones tomadas en cuenta para el cálculo (NP), y el año de inicio de las publicaciones en el tema en cuestión (*PY start*). La tabla fue ordenada según índice H decreciente y secundariamente con orden decreciente para TC, quedando en primer lugar la revista *Nucleic Acids Research*, con un valor de 120, seguida por *Database: The Journal of Biological Databases and Curation*, con un lejano valor de 25. El criterio inicial fue quedarse solo con las publicaciones en revistas cuyo índice H fuese mayor a uno, procurando un mínimo de dos publicaciones en el tema con por lo menos dos citas cada una. así es que al inicio de la tabla pueden verse revistas con altos valores para los tres índices, un alto número de citas y gran cantidad de publicaciones. Del total de 856 revistas en tabla, 203 habrían cumplido el criterio de corte, pero al acercarse a la división entre las revistas con criterio cumplido (H > 1, **en verde**) versus las de criterio no cumplido (H = 1, **en rojo**), saltan a la vista las siguientes interpretaciones:

- Por sobre la línea divisoria (H = 2), se hallan publicaciones con bajo número de citas, incluso en ciertos casos en revistas que llevan más de 10 años en el tema, versus otras que también con bajo número, lo lograron en muy pocos años. Ver revistas *Genes & Genomics* (posición 202) y *Medical Science Monitor* (posición 203)
- Bajo la línea divisoria (H = 1), se hallan publicaciones con un gran número de citas, en algunos casos mucho mayor a revistas con índice H igual a 2, incluso, más citas que revistas con índice H mayor a 10. Ver revistas *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* (204) y *Scientific Reports* (8)

- La cantidad total de citas, o por lo menos, el de las publicaciones más citadas, sí lo considera el índice G, pero este no muestra variación en las revistas cercanas a la línea divisoria. Ver revistas de las posiciones 196 a la 211.
- Las revistas *Genes & Genomics* (202) *Y Medical Science Monitor* (203), ambas con H = 2, llevan publicando en el tema 10 años y 2, respectivamente. Al compararlas solo por el índice H, no se aprecia el estancamiento en el tiempo de la primera, o al contrario, del rápido avance de la segunda, es el cociente *m* el que sí da cuenta de ello, *m* = 0,18 y *m* = 0,67 para la segunda.

Es por ello, que, al considerar solo el índice H de la revista, se dejaría fuera el factor citas totales y también fuera el factor tiempo. Se decidió por ello considerar además del índice H, el número total de citas de la revista, y el tiempo que lleva la revista publicando en el tema, esto último para valuar a las revistas más recientes de una forma más equilibrada con respecto a las de mayor trayectoria, al no exigirles el mismo impacto que el que podría haber alcanzado una con varios años más, y a su vez, exigiéndole más a una revista que lleva varios años.

Tabla 9. '*Source Impact'*. Impacto de las revistas según su índice H y otros indicadores internos (extracto de la tabla completa).

Posición	Revista	Índice H	Índice G	Cociente *m*	TC*	NP**	Año inicial
1	*Nucleic Acids Research*	**120**	**317**	**10,91**	**101379**	**406**	**2012**
2	*Database: The Journal of Biological Databases and Curation*	**25**	43	2,27	2619	153	2012
3	*PLOS ONE*	**23**	44	2,09	2132	69	2012
4	*Bioinformatics*	**19**	47	1,73	2233	56	2012
5	*BMC Genomics*	**16**	28	1,33	840	34	2011
6	*BMC Bioinformatics*	**15**	24	1,36	693	42	2012
7	*Plant and Cell Physiology*	**14**	21	1,27	1102	21	2012
8	*Scientific Reports*	**12**	26	1,20	743	26	2013
⋮							
196	*Journal of Database Management*	**2**	2	0,40	7	2	2018
197	*Journal of the American Water Resources Association*	**2**	2	0,25	7	2	2015
198	*Neuroinformatics*	**2**	2	0,18	7	2	2012
199	*Neurocomputing*	**2**	2	0,29	6	2	2016
200	*Neural Regeneration Research*	**2**	2	0,18	6	2	2012
201	*Electronic Library*	**2**	2	0,25	5	2	2015
202	*Genes & Genomics*	**2**	2	0,18	5	2	2012
203	*Medical Science Monitor*	**2**	2	0,67	4	2	2020
204	*Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*	**1**	1	0,14	5360	1	2016

Tabla 9. '*Source Impact'*. Impacto de las revistas según su índice H y otros indicadores internos (extracto de la tabla completa).

Posición	Revista	Índice H	Índice G	Cociente *m*	TC*	NP**	Año inicial
205	*Journal of Atmospheric and Oceanic Technology*	1	1	0,09	944	1	2012
206	*Physical Chemistry Chemical Physics*	1	1	0,17	688	1	2017
207	*Human and Ecological Risk Assessment*	1	1	0,14	563	1	2016
208	*Fungal Diversity*	1	1	0,13	522	1	2015
209	*Molecular Biology and Evolution*	1	1	0,10	456	1	2013
210	*Systematic Reviews*	1	1	0,17	378	1	2017
211	*Solar Energy*	1	1	0,09	316	1	2012
⋮							
849	*Molecular Endocrinology*	1	1	0,10	1	1	2013
850	*Advances in Electrical and Computer Engineering*	1	1	0,09	1	1	2012
851	*Animal Cells and Systems*	1	1	0,09	1	1	2012
852	*Interdisciplinary Sciences: Computational Life Sciences*	1	1	0,09	1	1	2012
853	*International Journal on Artificial Intelligence Tools*	1	1	0,09	1	1	2012
854	*The Journal of the Balkan Tribological Association*	1	1	0,09	1	1	2012
855	*South African Journal of Science*	1	1	0,09	1	1	2012
856	*Willdenowia*	1	1	0,08	1	1	2011

* Total de citas

** Cantidad de publicaciones, solo incluye aquellas con por lo menos una cita.

7.3.2. Creación de un nuevo índice para la evaluación del impacto de revistas a través del tiempo.

La dependencia del índice H del tiempo, hace de él una mala herramienta para evaluar a investigadores jóvenes, o en este caso, a revistas recientes, esto ya lo comentó el mismo Hirsch (2005) al proponer su indicador. y ha sido ampliamente discutido por otros investigadores (ver punto 3.2.2.2.3). Así pues, considerar únicamente el número de citas como parámetro para medir el impacto, también sería fuertemente influenciado por el tiempo —véase punto 3.2.2.2—. La dependencia del tiempo del índice H es resuelta por el cociente *m*, pero este no considera el total de citas, solo es una variante anualizada del índice H. En la Tabla 9 puede verse desde la línea 204 hacia abajo, donde comienzan las revistas con índice H igual a 1, una gran cantidad de citas en orden decreciente no hace variar ni al índice H ni al índice G, solo el cociente *m* registra variación, pero no en un orden decreciente como las citas. Al ver la variación del cociente *m* en las filas 196, 197 y 198, revistas con igual índice H y e igual cantidad de citas, el cociente *m* es mayor cuando más joven es la revista, y menor cuando es una revista más antigua.

Tras este análisis surgen las siguientes preguntas: ¿Puede aplicarse un barrido por impacto basado en el índice H de las revistas científicas sobre un *dataset* bibliográfico referente a un tema de investigación particular, que permita mejorar sus métricas confiriéndole una mejor calidad para ser posteriormente ocupado en una revisión sistemática? ¿Puede aplicarse un barrido por impacto basado en el índice H de las revistas científicas sobre un *dataset* bibliográfico referente a un tema de investigación particular, que permita mejorar sus métricas confiriéndole una mejor calidad para ser posteriormente ocupado en una revisión sistemática? ¿Es posible conseguirlo sin que esto signifique perder registros valiosos por su impacto si se consideran otros parámetros como la cantidad de citas per sé?

7.3.3. Definiciones para el nuevo índice de impacto.

Para valorizar el impacto de una revista en un tema determinado, tomando en cuenta el paso del tiempo y a la vez publicaciones únicas (en el tema) que de todas formas sí han conseguido una gran cantidad de citas, se propone utilizar el índice H interno, añadiendo el total de las citas de la revista, pero con un carácter anualizado.

Siendo la variación en el tiempo el transcurso entre que la revista haya comenzado a publicar en el tema y el año actual, tomando en cuenta ambos años, matemáticamente se formula como un Δt+1, siendo,

$$\Delta t = A\tilde{n}o\ en\ curso\ (2022) - a\tilde{n}o\ de\ inicio\ de\ las\ publicaciones$$

para el total de citas anualizado se define:

$$Total\ de\ citas\ anualizado\ (TCA) = \frac{TC}{\Delta t + 1} \qquad \textbf{(1)}$$

Se propone entonces un nuevo índice, en función del original índice propuesto por Hirsch, a nombrar provisionalmente como F_H, resultando de la multiplicación del índice H y el TCA, de la forma:

$$índice\ F_H = índice\ H \cdot \frac{Total\ de\ citas}{\Delta t + 1}$$

ó

$$índice\ F_H = índice\ H \cdot TCA \quad \textbf{(2)}$$

Por otra parte, siendo el cociente *m* una anualización del índice H, (Bornmann et al., 2008; Hirsch, 2005):

$$cociente\ m = \frac{índice\ H}{años} \quad \textbf{(3)}$$

el índice F_H también es equivalente a:

$$F_H = m \cdot TC \quad \textbf{(4)}$$

7.3.4. Aplicación del nuevo índice de impacto F_H.

Con el objetivo de aplicar el nuevo índice de impacto al *dataset* post deduplicación, se partió de la misma tabla *Source Impact*, esta vez, incorporando dos columnas a la tabla original, para el TCA y el índice F_H, quedando de la siguiente manera:

	A	B	C	D	E	F	G	H	I
1	Journal	F_H_index	TCA	h_index	g_index	quotient_*m*	TC	NP	PY_start
2	NUCLEIC ACIDS RESEARCH	1105952,7	9216	120	317	10,91	101379	406	2012

Figura 8. Nuevas columnas en tabla Sources Impact. Encabezado y primera fila.

Cada celda de la columna **C** correspondiente al TCA lleva la misma función —según la ecuación **(2)**—, por ejemplo, la celda C**2** lleva la función =G**2**/(2022-I**2**+1). Cada celda de la columna **B** correspondiente al F_H, lleva la misma función —según la ecuación **(1)**—, por ejemplo, la celda B**1** = C**2***D**2**.

Al aplicar el nuevo índice, un gran número registros que habrían quedado fuera al considerar solo el índice H, sí fueron consideradas. (Ver Tabla 10)

Tabla 10. *Source Impact* en orden decreciente según F_H, incluye TCA.

Posición	Revista	Índice F_H	TCA	Índice H	Índice G	*Cociente m*	TC	NP	Año inicial
1	Nucleic Acids Research	**1105952,7**	9216	120	317	10,91	101379	406	2012
2	*Database: The Journal of Biologic…*	**5952,3**	238,1	25	43	2,27	**2619**	153	2012
3	PLOS ONE	**4457,8**	193,8	23	44	2,09	**2132**	69	2012
⋮									
10	Journal of Cheminformatics	**777,6**	64,8	12	19	1,20	**648**	19	2013
11	*Acta Crystallographica Section B:…*	**765,7**	765,7	1	1	0,14	**5360**	1	2016
12	Human Mutation	**677,1**	84,6	8	12	0,73	**931**	12	2012
⋮									
27	Journal of Applied Crystallography	**128,1**	25,6	5	5	0,63	**205**	5	2015
28	Physical Chemistry Chemical Physi…	**114,7**	114,7	1	1	0,17	**688**	1	2017
29	Journal of Biotechnology	**112,0**	56,0	2	2	0,33	**336**	2	2017
⋮									
167	Journal of the American College o…	**9,8**	9,8	1	1	0,11	**88**	1	2014
168	Information Systems	**9,7**	4,8	2	2	0,33	**29**	2	2017
169	International Journal of Robotic R…	**9,5**	9,5	1	1	0,09	**105**	1	2012
⋮									
469	Annals of Thoracic Surgery	**2,0**	2,0	1	1	0,33	**6**	1	2020
470	Water Resources Research	**1,9**	1,9	1	1	0,11	**17**	1	2014
471	Drug Design Development and Th…	**1,9**	1,9	1	1	0,13	**15**	1	2015
⋮									
484	Urology	**1,8**	1,8	1	1	0,25	**7**	1	2019
485	Journal of the American Water Re…	**1,8**	0,9	2	2	0,25	**7**	2	2015
486	Anatomical Record-Advances in In…	**1,8**	1,8	1	1	0,13	**14**	1	2015
⋮									
828	Hydrological Processes	**0,2 (0,17)**	0,2 (0,17)	1	1	0,17	**1**	1	2017
⋮									
843	Crustaceana	**0,1 (0,11)**	0,1 (0,11)	1	1	0,11	**1**	1	2014
⋮									
852	Animal Cells and Systems	**0,1 (0,09)**	0,1 (0,09)	1	1	0,09	**1**	1	2012

7.3.5. Análisis inicial de la tabla *Source Impact* y su reordenamiento según el nuevo índice F_H.

El siguiente análisis es sobre la Tabla 10, excepto cuando se haga referencia a otra.

En las primeras tres posiciones logran permanecer en sus lugares las mismas revistas que encabezaban la Tabla 9, según el índice H. En este caso se trata de revistas con alto índice H y alto TCA.

En las posiciones 11 y 18, logran subir revistas que habían sido descartadas por su índice H (igual a 1), pero su alta cantidad de citas las sube hasta el primer ciento de las posiciones.

En las posiciones 167 y 169, dos revistas, al igual que en el caso anterior, suben de categoría (no aprobadas), hasta quedar bastante por sobre la línea aprobatoria gracias a su alta cantidad de citas (TC). Sin embargo, queda en la posición 167 una revista con 88 citas, y dos lugares más abajo una con 105 citas. Aquí se demuestra el carácter anualizado del índice, pues la revista de la posición 167 acumuló 88 citas desde el 2014, mientras que la del 169 acumuló las 105 desde el 2012, por ello es que se toma en cuenta no solo el TC, si no que el TCA.

En el grupo 469 a 471, justo en la nueva división de aprobadas y no aprobadas, las revistas no acumularon una cantidad de citas muy alta, dos de ellas, las primeras en quedar bajo la línea divisoria acumularon 17 y 15 citas, pero la que sí quedó por encima, con solo 6 citas aprobó, al haberlas conseguido en un período más corto (TCA).

Desde la posición 470 hacia abajo se hallan en general revistas con bajo índice H y TCA, pero en ciertos casos, como en la posición 485, encontramos lo que correspondería a un falso positivo según el criterio inicial, una revista que habría quedado por sobre la línea aprobatoria (índice H $\geq$ 2), pero en realidad cuenta con un bajo TCA.

Por último, este nuevo índice F_H —también vale para TCA— no tiene la limitación de mostrar solo valores enteros. En esta tabla se muestra solo 1 decimal, que incluso para el primer ciento de la tabla podría parecer innecesario, sin embargo, si se trabaja con una gran cantidad de datos, podría desearse considerar más decimales, y así poder ver pequeñas diferencias como se muestra en las posiciones 828, 843 y 852, donde tanto los valores del índice H y el TCA es bajo, de hecho en los tres casos el TC es 1, pero al ser de año de inicio distintos, se puede apreciar una leve diferencia en el valor final del índice.

7.3.6. Correcciones de datos posterior a la aplicación del nuevo índice F_H.

7.3.6.1. *Correcciones por año de inicio en tabla Source Impact.*

Al ya haber formulado el nuevo índice y ordenado la tabla *Source Impact* según dicha medición, diez líneas (la de posición 787 y las últimas nueve de la tabla) figuraban con un total de citas muy diferente al de las otras revistas de posiciones contiguas, su TCA era muy cercano

a cero, incluso en casos con más de diez o más de 100 citas. Las diez revistas señaladas tienen en común que no registraron el año de inicio de las publicaciones, ello explica por qué tampoco Biblioshiny pudo calcular correctamente el cociente *m*, columna sin datos también en estas diez filas. La columna F_H calcula de forma automática para todas las revistas el valor a través de la multiplicación de su índice H y el TCA, y este último dato depende de celdas cuyo valor figura en blanco, las del año de inicio.

Para corregir estos valores, se procede a revisar en el *dataset* (resultante del proceso de deduplicación), las revistas cuyos años de inicio de publicación figuran faltantes en la tabla *Source Impact*, resultando en que todas estas contaban con publicaciones indexadas en WoS bajo la categoría de "acceso temprano" (*Early Access*), por este motivo dichos registros no contaban con el metadato de la fecha de publicación (*Publication Year*), pero sí el de la fecha en que fueron liberadas bajo la categoría de acceso temprano (*Early Access Date*). Para cada una de estas revistas se revisó la publicación más antigua que cuente con al menos una cita[13], tomando en cuenta el campo *Publication Year* y *Early Access Date*, completando manualmente el campo año de inicio en la tabla *Source Impact*. Con esto las fórmulas de los campos F_H y TCA corrigen automáticamente el dato erróneo en la tabla.

El campo que no se corrige automáticamente es el del cociente *m*, pues corresponde a un dato cuyo cálculo lo realiza directamente Biblioshiny, tal como el índice H, por lo que se corrige agregando en estas filas la fórmula correspondiente para el cálculo del cociente *m* —ecuación **(3)**—, contando ya con todos los años de inicio. Este error solo se produjo en el caso de revistas que contaban con publicaciones *Early Access* efectivamente citadas. El resto de publicaciones *Early Access* del *dataset* no influyó de ninguna forma en alguno de los cálculos hechos por Biblioshiny para los indicadores mostrados en la tabla *Source Impact*.

La Tabla 11 muestra las correcciones hechas a raíz de los años de inicio faltantes. Y los cambios en las posiciones de las revistas. Las diez subieron de posición, pero solo cuatro de ellas quedaron por sobre el criterio de corte.

[13] Esto con el fin de mantener la fórmula original para el cálculo del índice H, y todos sus derivados, pues dichos indicadores basan su medición únicamente en las publicaciones efectivamente citadas.

Tabla 11. Errores y correcciones por año de inicio en tabla *Sources Impact*.

Filas con posición y datos erróneos pre corrección									
Posición	**Revista**	**Índice F_H**	**TCA**	**Índice H**	**Índice G**	***Cociente m***	**TC**	**NP**	**Año inicial**
787	Molecular Ecology Resources	**0,27**	0,07	4	5	—	139	5	—
849	Journal of Information Science	**0,03**	0,01	3	3	—	20	3	—
850	Glycobiology	**0,02**	0,01	2	3	—	23	3	—
851	European Journal of Human Genetics	**0,01**	0,01	1	2	—	11	2	—
852	Concurrency and Computation-Practice &…	**0,00**	0,00	1	2	—	6	3	—
853	Earth Science Informatics	**0,00**	0,00	1	2	—	5	**3**	—
854	International Journal of Social Research M…	**0,00**	0,00	1	1	—	4	1	—
855	Journal of Biomolecular Structure & Dina…	**0,00**	0,00	1	1	—	2	**1**	—
856	Amino Acids	**0,00**	0,00	1	1	—	1	**1**	—
857	Digestive Diseases and Sciences	**0,00**	0,00	1	1	—	1	1	—
Filas con posición y datos correctos post corrección									
Posición	**Revista**	**Índice F_H**	**TCA**	**Índice H**	**Índice G**	***Cociente m***	**TC**	**NP**	**Año inicial**
41	Molecular Ecology Resources	**69,50**	**17,38**	4	5	**0,50**	139	5	**2015**
106	Journal of Information Science	**20,00**	**6,67**	3	3	**1,00**	20	3	**2020**
248	Glycobiology	**5,75**	**2,88**	2	3	**0,25**	23	3	**2015**
462	European Journal of Human Genetics	**2,00**	**2,00**	1	1	**0,50**	4	1	**2021**
560	Concurrency and Computation-Practice &…	**1,22**	**1,22**	1	2	**0,11**	11	2	**2014**
594	Earth Science Informatics	**1,00**	**1,00**	1	2	**0,20**	5	4	**2018**
616	International Journal of Social Research M…	**1,00**	**1,00**	1	1	**1,00**	1	2	**2022**
643	Journal of Biomolecular Structure & Dina…	**0,86**	**0,86**	1	2	**0,14**	6	3	**2016**
677	Amino Acids	**0,67**	**0,67**	1	1	**0,33**	2	2	**2020**
727	Digestive Diseases and Sciences	**0,50**	**0,50**	1	1	**0,50**	1	1	**2021**

7.3.6.2. *Corrección de TC por publicaciones de igual título sin ser duplicados.*

En la sección 7.2.2 durante el proceso de deduplicación, se da cuenta de 34 registros que no contaban con DOI recuperable, lo que requirió de la búsqueda de duplicados aplicando formato condicional al campo título, y se descartó la presencia de registros de publicaciones en duplicado. Sin embargo, cuatro de los siete registros de publicaciones tituladas "*Database resources of the National Center for Biotechnology Information*", figuraban con la misma cantidad de citas (ver Tabla 12).

Tabla 12. Conteo erróneo de citas para publicaciones de la base de datos del NCBI

Año de publicación	Citaciones en WoS CC	Citaciones totales (TC)	Citaciones totales (TC) editado según fuente*
2012	2530	**2563**	**483**
2013	2530	**2563**	**410**
2014	286	293	—
2015	172	173	—
2016	228	231	—
2018	2530	**2563**	**909**
2019	2530	**2563**	**325**

* Datos proporcionados por la fuente de las publicaciones —*Nucleic Acids Research*— al 02-06-2022.

Estos datos fueron revisados en la web del publicante y efectivamente estaban erróneos, por lo que fueron corregidos en el *dataset*, pues no hacerlo implica un cálculo erróneo para los índices de impacto que dependen del valor de TC (ver Tabla 13). La posición de la revista en la tabla Source Impact sigue siendo la misma, al igual que los valores de índice H y cociente m, pues no dependen de TC, no así los índices G, F_H y el TCA, que sí cambian pues dependen de TC.

Tabla 13. Cambios en índices de impacto para la revista *Nucleic Acids Research* tras corrección de error en TC

Naturaleza de los datos	Posición	F_H	TCA	Índice H	Índice G	Cociente *m*	TC*	NP**	Año inicial
No corregidos	1	**1105952,7**	**9216**	120	**317**	10,91	**101379**	406	2012
Corregidos	1	**1019050,9**	**8492**	120	**304**	10,91	**93413**	406	2012

Este error en el conteo de citas es propio de WoS, cuyo sistema de conteo de duplicados se confunde al tratar con registros de igual título, sumando las citas de diferentes publicaciones de igual nombre, aun siendo artículos científicos diferentes.

7.3.7. Generación del nuevo archivo de entrada para Biblioshiny.

Al tener claros los parámetros de corte para el barrido por índice de impacto, y ya haberlo aplicado en la tabla *Source Impact*, resta aplicarlo al *dataset*, el archivo que contiene toda la información bibliográfica de los resultados de nuestra búsqueda bibliográfica, para con ello contar con un nuevo archivo de entrada para el análisis en Biblioshiny en base a los resultados de la deduplicación.

7.3.7.1. *Aplicación del barrido según índice F_H sobre archivo Excel del dataset deduplicado.*

Se agregó una columna adicional a la hoja de cálculo del *dataset* ya sin duplicados, incluyendo en cada celda de dicha columna una fórmula para el copiado automático en cada fila del índice F_H correspondiente a la revista donde fue publicado el artículo, extrayendo dicha información de otra hoja con la Tabla 10 ya corregida. Esta fórmula fue del tipo BUSCARV, como puede verse en la ecuación **(5)**, la que muestra como ejemplo la fórmula para el primer registro en la tabla:

$$= \text{BUSCARV}(\text{J2}; '\text{Source Impact}'! \$\text{A}\$2: \$\text{I}\$857; 2;) \qquad \textbf{(5)}$$

Ya con el dato del índice F_H en el *dataset*, este se ordenó según valores decrecientes de F_H —y de forma secundaria TC—. Esto dejó a 1951 registros por sobre la línea de corte —ya establecida previamente en la tabla *Source Impact*—, obteniendo con esto el número final de publicaciones a trabajar con el análisis bibliométrico posterior. Los 596 registros de publicaciones que quedaron por debajo de la línea de corte fueron borrados de la hoja de cálculo. Esta hoja de cálculo contiene todos los metadatos bibliográficos necesarios para nuestro análisis bibliométrico, ahora con el número final de publicaciones a analizar.

Ya habiendo generado el *dataset* que cumple con la deduplicación y el barrido por impacto, es posible utilizar sus datos para análisis varios usando el mismo Excel. Sin embargo, este archivo no sirve como archivo de entrada para Biblioshiny. De igual forma es conservado como insumo para posteriores análisis, pero se hace necesario replicar el barrido de impacto de forma tal que genere un archivo de entrada apto para Biblioshiny.

7.3.7.2. *Replicación del barrido según índice F_H sobre archivo Excel del dataset deduplicado en formato Bibliometrix Export File.*

El procedimiento descrito en el punto 8.3.7.1, es efectivo en cuanto logra el barrido por impacto sobre el *dataset*, lo deja en formato Excel, posibilitando un posterior análisis bajo la misma plataforma ofimática. Al no servir como archivo de entrada para Biblioshiny se explora la posibilidad de repetir el barrido con el método ya descrito, esta vez sobre un formato de archivo sí legible por Biblioshiny.

Anteriormente, al comienzo del segundo barrido, fue descrito el procedimiento de carga de archivos en Biblioshiny, seleccionando en el submenú *"Please choose want to do"* la opción *Import raw file(s)*. Esto implica que en el punto c) el archivo con el *dataset* deba ser subido en formato de texto (.txt). Después de cargado el archivo, fue generada la tabla *Source Impact*, descargada como Excel (.xlsx) para su modificación y utilización de los datos del índice de impacto calculado para el descarte de publicaciones en el *dataset*, también en formato Excel.

Ahora, para conseguir un archivo Excel por Biblioshiny, se efectuó el siguiente procedimiento:

a) se cargó el *dataset* deduplicado en Biblioshiny repitiendo los mismos pasos ya descritos en 7.3.1.1.
b) en el menú principal *Data/Load Data*, submenú *"Export Collection"*, en *Save as* se seleccionó la opción *Excel*, guardando este archivo de salida.

Con ello se genera un archivo Excel con el *dataset* deduplicado apto para ser leído posteriormente por Biblioshiny. El barrido de impacto se realiza sobre este archivo Excel, siguiendo lo descrito en el punto 8.3.7.1, pero reduciendo los cambios en el archivo al mínimo, para que la lectura posterior del mismo por parte de Biblioshiny resulte sin problemas, ya que este formato de archivo, viene con un orden específico de columnas y encabezados de las mismas propio. Estos pasos resumen lo descrito:

c) en Excel, se añadió en una hoja nueva una copia de la hoja Source Impact que incluye los datos del índice F_H
d) en la hoja del *dataset* (*Sheet 1*) se añadió una columna justo después de la contendora del nombre completo de las fuentes —con el encabezado SO—
e) ya en la columna recién agregada, en la fila 1 agregar encabezado F_H
f) en la fila 2 —la primera fila contiene los encabezados— se incluyó la fórmula BUSCARV ya descrita en 8.3.7.1 —ecuación **(5)**—, copiándola además hasta la última fila con datos (2548)
g) en el menú *Datos/* se ordenó la hoja completa según los valores de la columna F_H en orden decreciente

h) se eliminó: todas las filas con índice F_H < 2 —cuidando de no eliminar la fila de encabezados—, la columna F_H y la hoja Source Impact
i) se guardó el archivo.

Tras estos cambios, se obtuvo un archivo Excel con el *dataset* post barrido de impacto apto para ser sometido a análisis por Biblioshiny.

7.4. Bibliometría sobre *dataset* depurado

Los puntos 7.2 y 7.3, correspondientes a barrido por deduplicación y barrido por impacto respectivamente, logran eliminar errores producto de la duplicación de registros y enfocarnos en los registros de artículos científicos que sí han generado un impacto, respectivamente. Esto obedece a los objetivos iniciales planteados para esta investigación. Los siguientes análisis serán hechos a partir de este proceso de depuración (puntos 8.2 + 8.3), a no ser que se especifique lo contrario.

7.4.1. Carga de archivos ya depurados en Biblioshiny.

En Biblioshiny, se cargó como archivo de entrada el *dataset* ya depurado, seleccionando:

a) menú principal *Data/Load Data*
b) en el submenú *Import or Load Files/Please choose want to do*, se especificó el tipo de archivo *Load Bibliometrix File(s)*
c) en el submenú *Choose a file/Browse*, se cargó el archivo Excel producto del proceso descrito en 8.3.7.2.
d) clic en "*Start*".

7.4.2. Análisis general y comparativo del *dataset* depurado.

En Biblioshiny, desde el menú *Overview/Main Information*, se obtuvo el siguiente gráfico con la información principal de nuestra base de datos, el *dataset* ya depurado. (Figura 9)

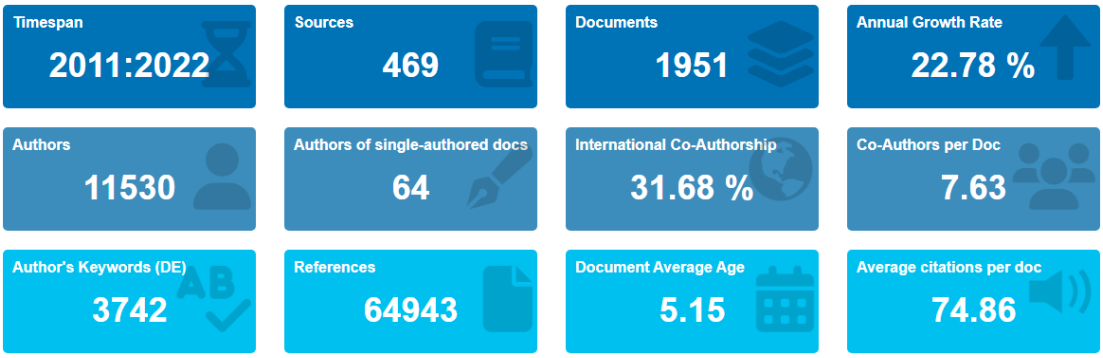

Figura 9. Información principal

¿Puede realizarse un análisis bibliométrico sobre él sin haber perdido información valiosa? Para un análisis mayor, en profundidad y responder de paso esta pregunta se muestra en la Tabla 14 un análisis comparativo con los datos previos al proceso de barrido por impacto. Esta corresponde a la tabla completa obtenida desde el mismo menú, a la que se adicionaron los resultados prebarrido de impacto, y el porcentaje de variación al aplicar este barrido.

Tabla 14. Información principal ampliada y comparativa

Descripción	Resultados		Variación (%)
	Pre-barrido	Post-barrido	
INFORMACIÓN PRINCIPAL			
Fuentes (Revistas, Libros, etc.)	980	469	-52,14
Documentos (publicaciones)	2547	1951	**-23,40**
Tasa anual de crecimiento %	23,87	22,78	-4,57
Edad promedio de las publicaciones	5,05	5,15	1,98
Promedio de citaciones por documento	58,15	74,86	**28,74**
Referencias	82689	64943	-21,46
CONTENIDO DE LOS DOCUMENTOS			
Keywords Plus (ID)	5755	4775	-17,03
Author's Keywords (DE)	5589	3742	-33,05
AUTORES			
Total de autores	14042	11530	-17,89
Autores de documentos de un solo autor	110	64	**-41,82**
COLABORACIÓN DE AUTORES			
Documentos de un solo autor	112	65	**-41,96**
Promedio de autores por documento	7,04	7,63	**8,38**
Co-autoría internacional %	30,19	31,68	**4,94**
TIPOS DE DOCUMENTO			
Artículo	2270	1761	-22,42
Artículo; Capítulo de libro	4	1	-75,00
Artículo; Data paper	48	45	-6,25
Artículo; Acceso temprano	30	6	**-80,00**
Artículo; Acta de congreso	56	41	-26,79
Review	134	97	-27,61
Review de bases de datos	1	0	-100,00
Review; Capítulo de libro	1	0	-100,00
Review; Acceso temprano	3	0	-100,00

7.4.2.1. *Intervalo de tiempo.*

Primero, el intervalo de tiempo (*Timespan*), coincide parcialmente con el intervalo especificado en la búsqueda sistemática realizada en WoS (punto 7.1), donde se especificó

Timespan: 2012-01-01 to 2022-05-16 (Index Date)

Esta pequeña diferencia entre 2011 o 2012 como año de inicio, radica en que no se tomó en las especificaciones de búsqueda la fecha de publicación, si no que la fecha de indexación a la base de datos de WoS (*Index Date*) por eso existen artículos con fecha de publicación anterior al 2012, pero con fecha de indexación 2011. Esta diferencia suele ser tan solo de un par de semanas, por lo que no implica grandes diferencias en los resultados obtenidos.

7.4.2.2. *Revistas y publicaciones.*

Lo primero que se destaca, es que las fuentes de los artículos (revistas principalmente), bajaron a la mitad, de 980 a 469 (reducción del 52,14%), sin embargo, los documentos analizados (los artículos científicos) bajaron de 2047 a 1951, reduciéndose solo un 23,40%. Para las publicaciones en general, sin distinción de fuentes, la tasa de crecimiento anual es de un 22,78%, con distinción por año visible en el Gráfico 1, obtenido desde el menú *Overview/Annual Scientific Production*.

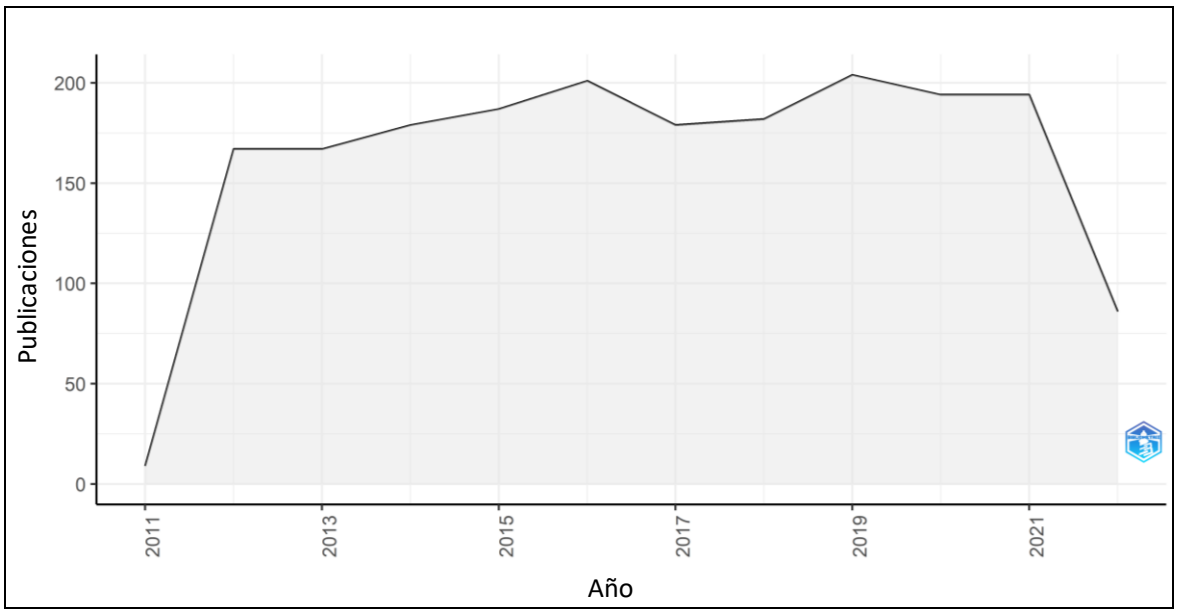

Gráfico 1. Producción científica anual

7.4.2.3. *Autores y colaboración.*

Con respecto a los autores, con el barrido de impacto hubo una reducción de un total de 14042 a 11530 (reducción del 17,89%). Sin embargo, los autores de documentos de un solo autor bajaron de 110 a solo 64 (reducción del 41,82%), símil con la reducción de 112 a 654 (reducción del 41,96%) para los documentos de un solo autor. Estas grandes variaciones porcentuales se interpretan como positivas, al ver junto con ello dos variaciones que —partiendo de una reducción del 23,40% del total de documentos (parámetro base para la interpretación del resto de las variaciones)— son de aumento, estas son el promedio de coautores por documento, pasando de un 7,04 a un 7,63 (aumento del 8,38%), y el porcentaje de coautoría internacional, pasando de un 30,19% a un 31,68% (aumento del 4,94%).

7.4.2.4. *Palabras clave (Keywords).*

Con respecto a las palabras clave, la cantidad total de *Keywords Plus* disminuyó un 17,03%, pero la de *Keywords* del autor, lo hizo en un 33,05%. De esto se interpreta que las *Keywords Plus* resisten (en cuanto a la variación de su cantidad total en el la base de datos analizada) casi el doble con respecto a lo que se reducen las *Keywords* del autor.

7.4.2.5. *Edad de las publicaciones y cantidad de referencias.*

La edad promedio de las publicaciones pasa 5,05 a 5,15 años, mostrando un ligero aumento.

La cantidad total de referencias registradas en el *dataset* pasa de 82689 a 64943, registrando una disminución del 21,5 %, casi a la par con el de la disminución del número de publicaciones (23,4%).

7.4.2.6. *Tipos de documento (Tipos de publicaciones).*

Para este apartado se incluye una ampliación de la tabla anterior (**Tabla 14**) añadiendo los porcentajes del tipo de documento con respecto al total antes y después del barrido, y las variaciones de estos porcentajes:

Tabla 15. Tipos de documento y variación post barrido

Tipos de documento	Pre-barrido		Post-barrido		Variación (%)	
	Cantidad	%	Cantidad	%	De las cantidades	De los porcentajes*
Todos (total de publicaciones)	**2547**	***100,00***	**1951**	***100,00***	**-23,40**	***0,00***
Artículo	2270	*89,12*	1761	*90,26*	-22,42	*1,28*
Review	134	*5,26*	97	*4,97*	-27,61	*-5,50*
Artículo; Data paper	48	*1,88*	45	*2,31*	-6,25	***22,39***
Artículo; Acta de congreso	56	*2,20*	41	*2,10*	-26,79	*-4,42*
Artículo; Acceso temprano	30	*1,18*	6	*0,31*	-80,00	*-73,89*
Artículo; Capítulo de libro	4	*0,16*	1	*0,05*	-75,00	*-67,36*
Review; Acceso temprano	3	*0,12*	0	*0,00*	-100,00	*-100,00*
Review de bases de datos	1	*0,04*	0	*0,00*	-100,00	*-100,00*
Review; Capítulo de libro	1	*0,04*	0	*0,00*	-100,00	*-100,00*

* Variación del porcentaje de tipo de archivo, tomando como 100% el porcentaje de tipo de archivo antes del barrido

En el resto de esta tesis se suele mencionar de forma general a las publicaciones como artículos, artículos científicos, o publicaciones, refiriéndose a lo mismo, como se entiende también el anglicismo *papers*. En esta tabla y en el desglose que le sigue con los siguientes puntos se usa el concepto de documentos o publicaciones para referirse al conjunto de estos, y solo se usa artículos para referirse a las publicaciones que sí están catalogadas bajo este concepto en lo que se refiere a tipo de documento o tipo de publicación.

7.4.2.6.1. *Artículos.*

De las 1951 publicaciones analizadas, los artículos son 1761, y corresponden a un 90,26% del total. Con el barrido pasaron de 2270 a 1971 (reducción del 22,42%), disminuyendo prácticamente en la misma razón que la totalidad de la muestra (23,40%), de hecho, el porcentaje de publicaciones tipo artículo aumenta un 1,28%, siendo este tipo de documento el menos afectado porcentualmente con el barrido, al tomar como punto de referencia el porcentaje del tipo de publicación antes del barrido.

7.4.2.6.2. *Reviews.*

El segundo tipo de publicación mayoritario fue el *Review*, con 97 del total, lo que corresponde a un 4,97% del total de la muestra. Este porcentaje de tipo de publicación disminuyó un 5,50% con respecto al porcentaje antes del barrido.

7.4.2.6.3. *Data paper.*

Los tipos de publicación desde la tercera fila hacia abajo, están clasificados en WoS en más de un tipo, esto vale solo para nuestra muestra. En este caso, las publicaciones clasificadas como *data paper*, además clasifican como artículo. Estos registros corresponden a 45 de 1951, lo que equivale a un 2,31% del total. Tras el barrido solo se redujo su cantidad de 48 a 45, lo que, si bien corresponde a una reducción del 6,25% de las publicaciones de este tipo, y comparando los porcentajes del tipo de publicación antes (1,88%) y del barrido (2,31%), por resta simple de ellos se entiende un aumento del 0,43%, en realidad representa un aumento proporcional del 22,39%, si se considera como 100% el porcentaje de este tipo de publicación antes del barrido. Es por esto que las variaciones de los porcentajes de los diferentes tipos de publicación fueron calculadas de esta forma y no por una resta simple.

7.4.2.6.4. *Actas de congreso (proceedings paper).*

Las publicaciones que clasifican en este tipo de publicación, clasifican también como artículo, es por ello que están en los resultados analizados, pues en las especificaciones de la búsqueda sistemática realizada en WoS no se indicó este tipo de publicación (solo quedaron por su doble clasificación). Estos registros corresponden a 41 de 1951, lo que equivale a un 2,10% del total. Tras el barrido se redujo su cantidad de 56 a 41, siendo esto una reducción del 26,79% de los artículos de este tipo, y una disminución del 4,42% en proporción al porcentaje de este tipo de archivo antes del barrido.

7.4.2.6.5. *Artículos bajo el concepto de acceso temprano (Early Access).*

Estas publicaciones *Early Access* son artículos, que no cuentan con todos los procesos de revisión por pares terminados, pero sí con una revisión y aprobación inicial, que los faculta para ser liberados tempranamente indicando su estado temporal.

Estos registros corresponden a seis de 1951, lo que representa solo el 0,31% del total. Tras el barrido disminuyeron de 30 a seis, lo que corresponde a una reducción de un 80,00%, y un 73,89% con respecto al porcentaje de este tipo de archivo antes del barrido.

7.4.2.6.6. *Capítulos de libro.*

Del total de registros analizados por Biblioshiny tras el proceso de barrido, es solo uno el que calza con la categoría de capítulo de libro —clasifica además como artículo—,

representando un 0,05% del total de la muestra. Al igual que el caso de las actas de congreso, queda en los resultados de búsqueda solo por su doble clasificación, pues no es un tipo de publicación que se haya especificado en los parámetros de la búsqueda sistemática. Tras el barrido, este tipo de publicación disminuye de cuatro a una, lo que corresponde a una disminución del 75,00%, y un 67,36% con respecto al porcentaje de publicaciones de este tipo antes del barrido. Este alto porcentaje de disminución se calcula sobre la base de un número muy bajo de publicaciones —solo cuatro—, por lo que se considera un dato de poca representatividad para este tipo de publicación.

7.4.2.6.7. *Review (de acceso temprano), Review de bases de datos y Review-Capítulo de libro.*

Los registros de publicaciones con estas clasificaciones, solo estaban presentes antes del barrido, por lo que disminuyeron (para ambos porcentajes de variación calculados) en un 100%. Sin embargo, antes del barrido solo eran tres, uno y uno, respectivamente, lo que equivale sumando los tres casos a un 0,16% del total pre barrido, mismo porcentaje para los capítulos de libro-artículo antes del barrido, y así como en tal caso, el gran porcentaje de disminución se considera un dato de poca representatividad para estos tipos de publicaciones.

7.4.2.7. *Promedio de citaciones por publicación*

7.4.2.7.1. *Promedio de citaciones por publicación para el total de la muestra analizada antes y después del barrido por impacto.*

El promedio de citaciones por publicación pasó de 58,15 a 74,86, lo que corresponde a un aumento del 28,74%. Este dato es de gran importancia, ya que el objetivo del barrido por impacto fue seleccionar para el análisis bibliométrico posterior a publicaciones que tuviesen un mínimo de impacto según los parámetros ya descritos en la generación del índice y el criterio de corte aplicado.

7.4.2.7.2. *Promedios de citaciones por publicación para cada año*

La siguiente Tabla 16 y el Gráfico 3 fueron obtenidos desde el menú *Overview/Average Citation Per Year*, opciones *Table* y *Plot* respectivamente. La tabla original muestra los promedios de citaciones por publicación de cada año para la base de datos post barrido y su equivalente anualizado; a estos dos fue añadida una columna con el porcentaje de variación del

promedio de citaciones por publicación para cada año con respecto al mismo valor del año anterior.

Tabla 16. Promedio de citas por publicación para cada año y promedio de citas anualizado de cada publicación

Año	NP	Promedio de citas por publicación*	Variación %**	Promedio de citas por publicación anualizado***	Años citables
2011	9	27,78	—	2,53	11
2012	167	136,53	391,47	13,65	10
2013	167	**141,39**	**3,56**	15,71	9
2014	179	121,75	-13,89	15,22	8
2015	187	66,58	**-45,31**	9,51	7
2016	201	115,40	**73,33**	**19,23**	6
2017	179	78,25	-32,19	15,65	5
2018	182	68,21	-12,83	17,05	4
2019	204	49,20	-27,87	16,40	3
2020	194	19,02	**-61,34**	9,51	2
2021	194	8,63	**-54,63**	8,63	1
2022	86	1,38	-84,01	—	0

*Se calcula sumando todas las citas registradas a la fecha de la búsqueda inicial para los artículos de cada año, dividiendo este valor por el total de publicaciones.
**Variación porcentual del P.D.C.P.P. (tercera columna) de cada año con respecto a su año anterior.
***Corresponde al mismo valor del P.D.C.P.P. (tercera columna), pero dividido por los años citables.

Se entiende que las citaciones por artículo vayan disminuyendo al comparar publicaciones más más antiguas versus las más jóvenes—asumiendo por lo menos un mínimo de impacto, en el caso contrario, al menos no van a disminuir—, sin embargo, esto no se aprecia en toda la tabla.

El primer porcentaje de variación para el promedio de citaciones por artículo anual, correspondiente a la variación 2011-2012, registra un altísimo valor, de más del 390%, lo que no se analizará en mayor grado por partir de un año con un número muy pequeño de publicaciones, por lo que el dato pierde validez.

El siguiente porcentaje de variación, 2012-2013, registra un aumento para el promedio de citaciones por artículo de casi un 4%, lo que no se condice con la idea del descenso generalizado para las publicaciones conforme se analizan grupos de estas de edad más joven, como se aprecia además en el Gráfico 2. El año 2012, podría ser quien marque el punto más alto en cuanto el promedio de citaciones, por ser este cálculo en base a las citas acumuladas para los artículos de cada año, sin embargo, el descenso generalizado comienza el 2013, no el 2012.

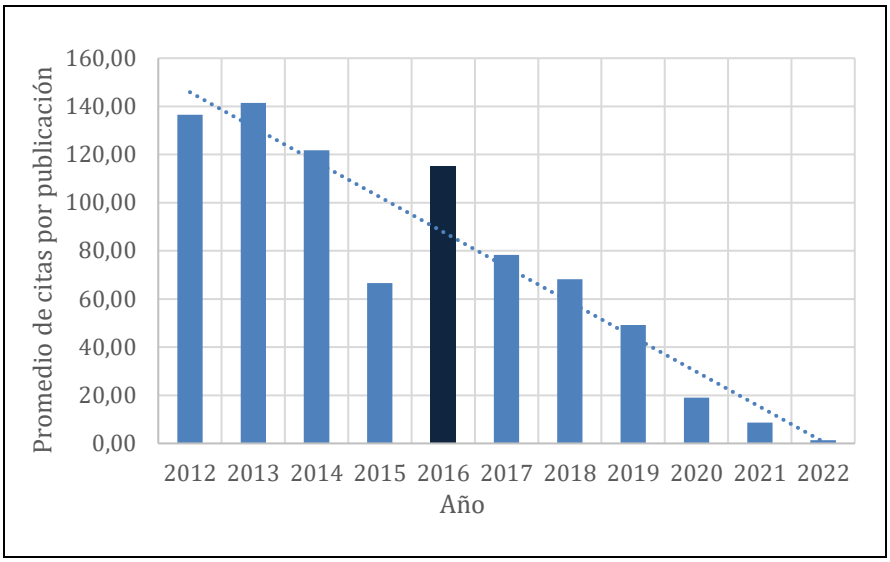

Gráfico 2. Línea de tendencia para el promedio de citas por publicación para cada año

Esta tendencia a un descenso generalizado en el promedio de citaciones por publicación posterior al 2013, registra una baja abrupta el año 2015, seguido de una fuerte alza el año siguiente, alza no despreciable solo por venir de una fuerte baja, pues al ver la línea de tendencia (Gráfico 2), el año que más sobrepasa la tendencia generalizada es precisamente el 2016.

Ambos años, 2013 y 2016, son los únicos que registran alzas con respecto a sus respectivos años anteriores, se infiere por ello la presencia de características específicas de los artículos publicados en esos años que le confieren a estos años tales alzas en el promedio de citaciones por artículo. Se infiere la presencia de unas pocas publicaciones con muy alta cantidad de citas —por fuera del promedio—, lo que le confiere estas alzas en el promedio de citaciones para los artículos de estos años.

Así como resaltan las alzas en los años ya mencionados, el período 2020-2021-2022 registra las más fuertes disminuciones en citaciones con respecto a sus años anteriores, incluyo según esta medida más fuertes que la brusca caída del 2015.

La quinta columna de la Tabla 16, corresponde al promedio de citaciones por artículo para cada año, anualizado. Ello se realiza para descontar la ventaja de los años para las publicaciones más antiguas, igualando la cancha para comparar el promedio de citaciones entre grupos de estas con amplias diferencias de edad —de la misma forma que Hirsch planteó el

cociente *m* al proponer el índice H—. La información de esta columna aparece graficada en el Gráfico 3.

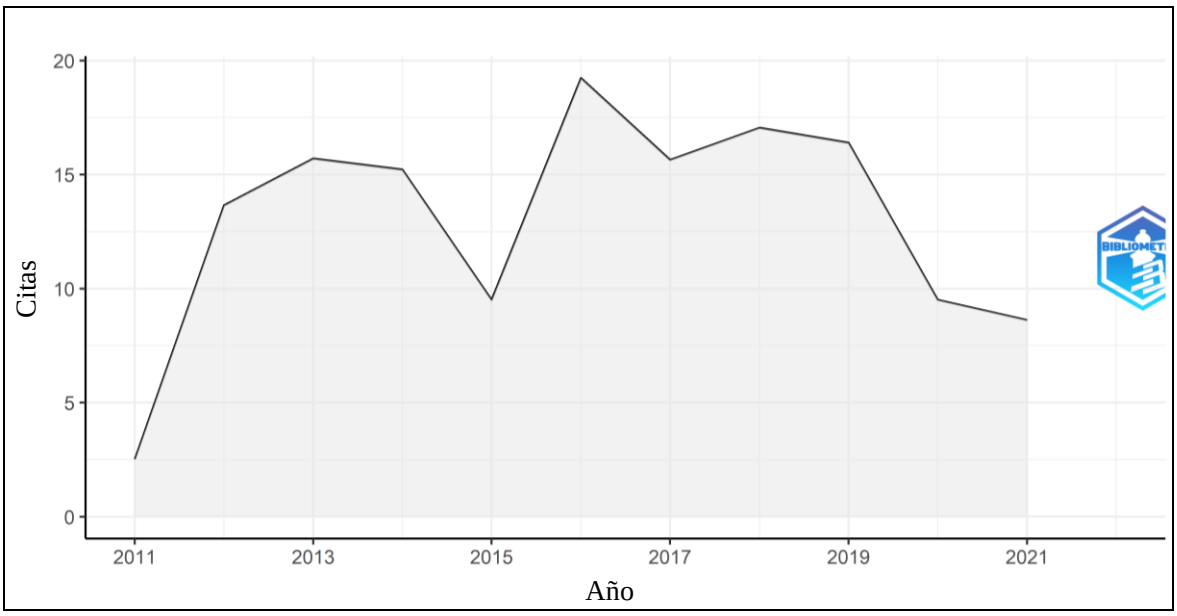

Gráfico 3. Promedio anual de citas por artículo

Aquí se puede ver claramente el año con el mayor promedio de citaciones anualizado, el 2016, seguido por el 2018. Ello indica que, los máximos valores para el promedio de citas por artículo para cada año, no muestran necesariamente los años cuyos artículos tienen más citas evaluándose con un carácter anualizado, tampoco lo indica el cálculo de la variación porcentual de este valor para cada año con respecto a su año anterior, es decir, al ver las variaciones 2017-2018 en las columnas 2 y 3 de la Tabla 16, solo se ven descensos, este segundo máximo del año 2018 solo se aprecia al darle un carácter anualizado a las citas.

7.4.3. Las fuentes.

7.4.3.1. *Tipos de fuentes presentes.*

Según los tipos de publicación presentes en el *dataset* (Tabla 15), las publicaciones de tipo artículo, que representan el 90,26% de la muestra, más los reviews, data papers, actas de congreso y artículos de acceso temprano, juntos suman un 99,95% de nuestro *dataset*. ¿Qué tienen en común estos cinco tipos de publicaciones?, ellos corresponden a publicaciones científicas provenientes de revistas científicas (*Journals*). El 0,05% restante, corresponde a una

sola publicación, un capítulo de libro, el que es parte de una serie de libros. Esta diferenciación puede revisarse además filtrando el archivo Excel del *dataset* por la columna *Publication type*[14], en donde aparecen solo dos categorías: J y S obteniéndose el siguiente resultado ahora por número de publicaciones (ver Tabla 17).

Tabla 17. Tipos de fuente

NP	Tipo
1950	(J) Revista
1	(S) Libro en serie

Es por esta amplia diferencia de 1950 a 1 en la muestra analizada en esta investigación, que, para esta sección referente a las fuentes, y en gran parte del documento completo de tesis, se ocupa el término revistas para referirse de forma genérica a todas las fuentes, no siendo necesariamente aplicable esta asociación a otros conjuntos de documentos científicos, pudiendo estos tener una diversidad de fuentes mucho más amplia.

7.4.3.2. *Revistas más relevantes.*

Una pregunta interesante al momento de evaluar la literatura científica sobre las bases de datos web, es qué revistas concentran la mayor cantidad de publicaciones en este tema —sin hacer aseveraciones de mayor profundidad como por ejemplo, en qué revistas los investigadores prefieren publicar sus trabajos sobre esta temática, pues para tal nivel de análisis se hace necesario profundizar en muchos otros aspectos relativos a las muchas variables que influyen en que el investigador publique en una u otra revista—. Según esta pregunta meramente cuantitativa, es que se definen las revistas más relevantes en el campo de las bases de datos web.

[14] El campo Publication Type se refiere a los tipos de registros bibliográficos según la naturaleza de publicación, esto en cuanto a quién la publica, en este caso haciendo la diferencia entre revistas o una editorial de libros. El campo que sí se refiere al tipo de documento, como se los diferenció en todo el punto 8.4.2.6Tipos de documento (Tipos de publicaciones). es Document Type.
Esta aclaración es válida para revisiones bibliométricas de *dataset*s provenientes de la base de datos Web of Science y su índice WoS-CC, y no necesariamente de otras bases de datos, pues cada una define sus criterios para la aplicación de filtros.

7.4.3.2.1. *Top 30 de las revistas más relevantes y distribución de Bradford.*

El Gráfico 4, obtenido desde el menú *Sources/Most Relevant Sources*, muestra las 30 revistas más relevantes.

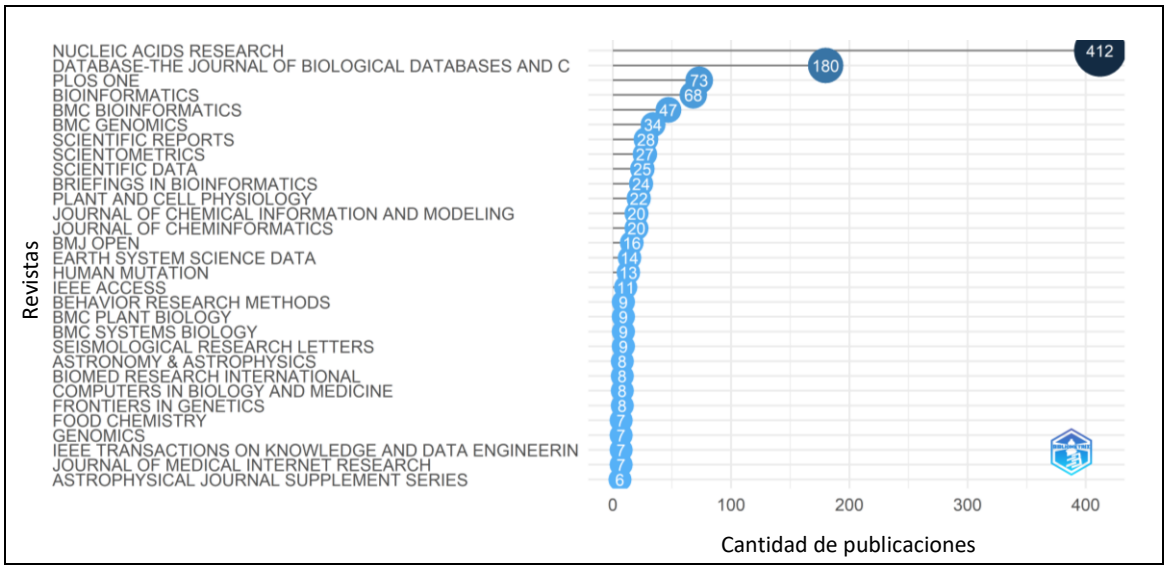

Gráfico 4. Revistas más relevantes (según cantidad de publicaciones en el tema)

Resalta con gran ventaja por sobre las demás la revista *Nucleic Acids Research* de la que provienen 412 publicaciones del total (un 21,12%), siendo por lejos la revista más relevante en cuanto a cantidad de publicaciones. La segunda más relevante es *Database: The Journal of Biological Databases and Curation*, con 180 publicaciones (un 9,23% del total), la que supera en más del doble la cantidad de publicaciones de la tercera en esta lista, *PLOS ONE*, con 73 publicaciones (un 3,74% del total). El ranking con distribución porcentual y editoriales se muestra en la Tabla 18.

Tabla 18. Las 30 revistas más relevantes, su editorial y distribución porcentual con respecto al total de la colección.

Posición.	Revistas	Editorial	NP	%
***1**	**Nucleic Acids Research**	**Oxford Univ Press**	**412**	**21,12**
***2**	**Database: The Journal of Biological Databases and Curation**	**Oxford Univ Press**	**180**	**9,23**
***3**	**PLOS ONE**	**Public Library Science**	**73**	**3,74**
4	**Bioinformatics**	**Oxford Univ Press**	**68**	**3,49**
5	BMC Bioinformatics	BMC	47	2,41
6	BMC Genomics	Biomed Central Ltd	34	1,74
7	Scientific Reports	Nature Publishing Group	28	1,44
8	Scientometrics	Springer	27	1,38
9	Scientific Data	Nature Publishing Group	25	1,28
10	Briefings in Bioinformatics	**Oxford Univ Press**	24	1,23

11	Plant And Cell Physiology	**Oxford Univ Press**	22	1,13
12	Journal of Cheminformatics	BMC	20	1,03
13	Journal of Chemical Information and Modeling	Amer Chemical Soc	20	1,03
14	BMJ Open	BMJ Publishing Group	16	0,82
15	Earth System Science Data	Copernicus Gesellschaft mbH	14	0,72
16	Human Mutation	Wiley	13	0,67
17	IEEE Access	IEEE-Inst Electrical Electronics Engineers Inc	11	0,56
18	BMC Plant Biology	BMC	9	0,46
19	Behavior Research Methods	Springer	9	0,46
20	BMC Systems Biology	BMC	9	0,46
21	Seismological Research Letters	Seismological Soc Amer	9	0,46
22	Astronomy & Astrophysics	EDP Sciences S A	8	0,41
23	Biomed Research International	Hindawi Ltd	8	0,41
24	Computers In Biology and Medicine	Pergamon-Elsevier Science Ltd	8	0,41
25	Frontiers In Genetics	Frontiers Media Sa	8	0,41
26	Journal of Medical Internet Research	JMIR Publications, Inc	7	0,36
27	Genomics	Academic Press Inc Elsevier Science	7	0,36
28	Food Chemistry	Elsevier Sci Ltd	7	0,36
29	IEEE Transactions on Knowledge and Data Engineering	IEEE Computer Soc	7	0,36
30	Astrophysical Journal Supplement Series	IOP Publishing Ltd	6	0,31
		Total de publicaciones en el top 30 de las revistas más relevantes: 1136		58,23%

** Las tres primeras revistas componen el núcleo de Bradford (ver también **Gráfico 5**)*

Estas primeras tres revistas de la lista concentran en total un 34,09%, aproximadamente un tercio de toda la colección. Esto coincide con la Ley de Bradford, (ver punto 3.2.2.1.1). Como se muestra en el Gráfico 5, obtenido desde *Sources/Bradford's Law*, el núcleo de Bradford, compuesto en este caso por tres revistas —las ya mencionadas— abarca un tercio de toda la colección.

En el gráfico se aprecia claramente una línea recta casi perfecta para el trazado de los primeros tres puntos, con un mínimo punto de inflexión en la segunda revista. Posterior al núcleo de Bradford, la línea continúa de forma más bien horizontal, con un decaimiento mínimo, para después de este tramo mostrar un decaimiento de tipo logarítmico, como lo plantea Bradford.

El comportamiento descrito por la curva en este pequeño tramo recién descrito, indica que la o las revistas que le siguen en el ranking de productividad en esta temática, resisten la fuerza del núcleo de Bradford mucho más que el resto de la colección. La revista que sigue del núcleo de Bradford, volviendo a mirar el Gráfico 5, se muestra con una cantidad de publicaciones mucho más cercana al tercer lugar que al quinto, contrario al decaimiento tipo logarítmico que describe el resto del gráfico. Este comportamiento anti-Bradford corresponde a la cuarta revista en el ranking, *Bioinformatics*.

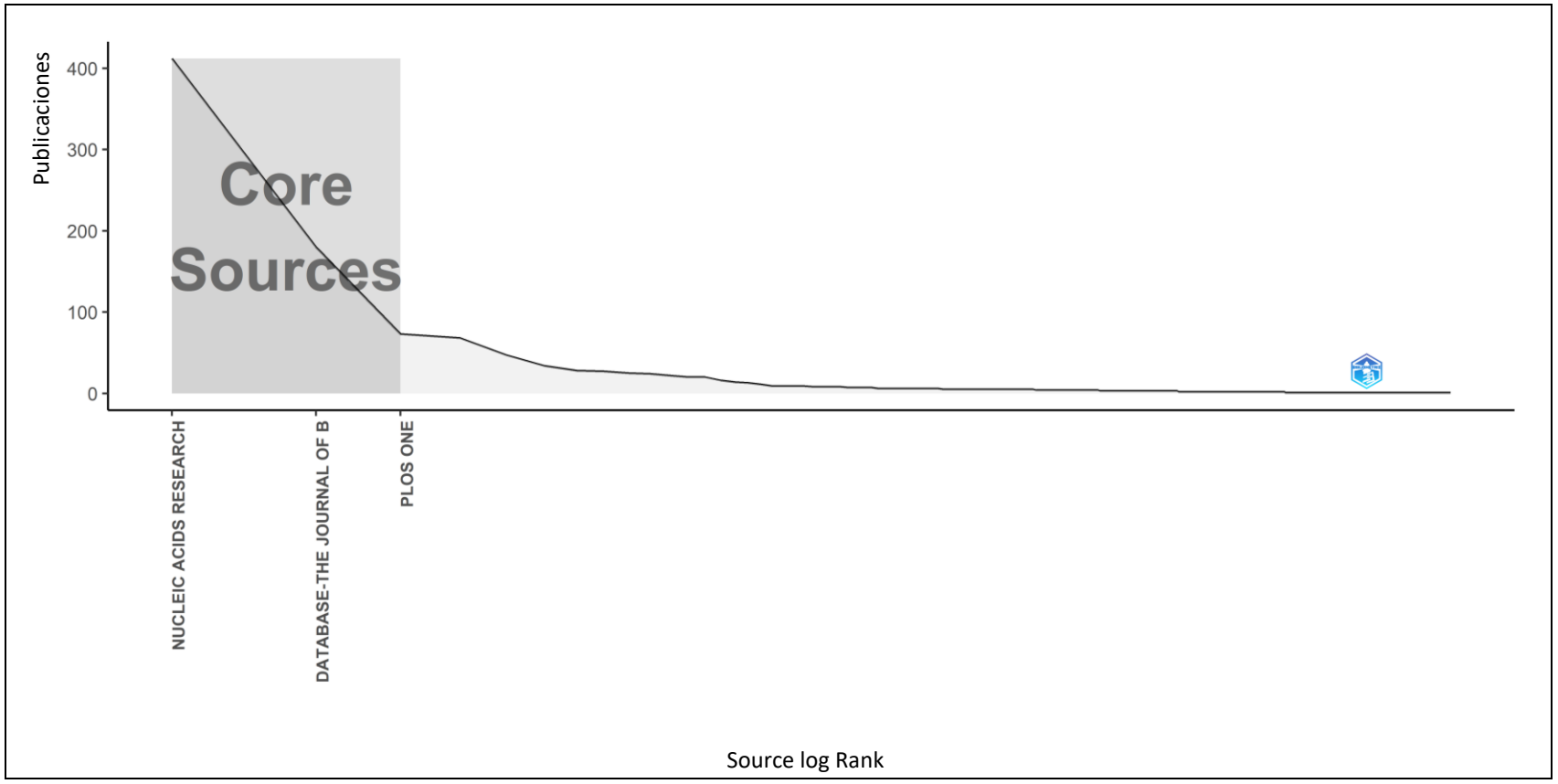

Gráfico 5. Distribución de Bradford. Señala el núcleo de las revistas donde los investigadores más publican sobre el tema. Este núcleo de tres revistas produce un tercio de las publicaciones en este tema.

7.4.3.2.2. *Núcleo de Bradford y la editorial de la Universidad de Oxford.*

La revista *Nucleic Acids Research,* es una de las más de 500 revistas de la Universidad de Oxford, de hecho, de los primeros diez puestos en este ranking (Tabla 18), cinco revistas son editadas por esta universidad:

Nucleic Acids Research
Database: The Journal of Biological Databases and Curation
Bioinformatics
Briefings in Bioinformatics
Plant And Cell Physiology

Así como se mostró la distribución y el núcleo de Bradford, con las tres revistas que equivalen a un tercio del total de la colección, específicamente el 34,09%, resulta interesante mencionar el caso de la concentración porcentual del total de revistas solo de la editorial de la Universidad de Oxford, pues sumados los porcentajes de las publicaciones de estas cinco revistas en la colección, equivalen a un 36,19%. Ello indica que dicha editorial, produce por sí sola —a través de las revistas mencionadas—, el equivalente al núcleo de Bradford completo e incluso más.

7.4.3.3. ***Relevancia a través del tiempo.***

En el caso del Gráfico 4, la Tabla 18 y Gráfico 5, correspondientes a los ya expuestos en 8.4.3.2.1 y 8.4.3.2.2, la variable medida fue la cantidad de publicaciones para cada revista, contando para ello todas las emanadas dentro del intervalo de tiempo evaluado en esta investigación, por ello coinciden en los valores máximos (412 publicaciones para el caso de la revista más relevante). Un aspecto aun no discutido es que estos valores son un recuento de su productividad año tras año, contando cada revista con el peso —para bien o para mal— de su desempeño años anteriores para posicionarse hoy en el lugar que hoy les merece.

Para evaluar su desempeño a través del tiempo, se grafica la productividad año a año de forma acumulada desde el menú *Sources/Source Dynamics*, con el parámetro *Occurrences* opción *Cumulate* (Gráfico 6), y luego sin acumulación, mostrando la productividad de cada año de forma independiente, cambiando a la opción *Per Year* (Gráfico 7).

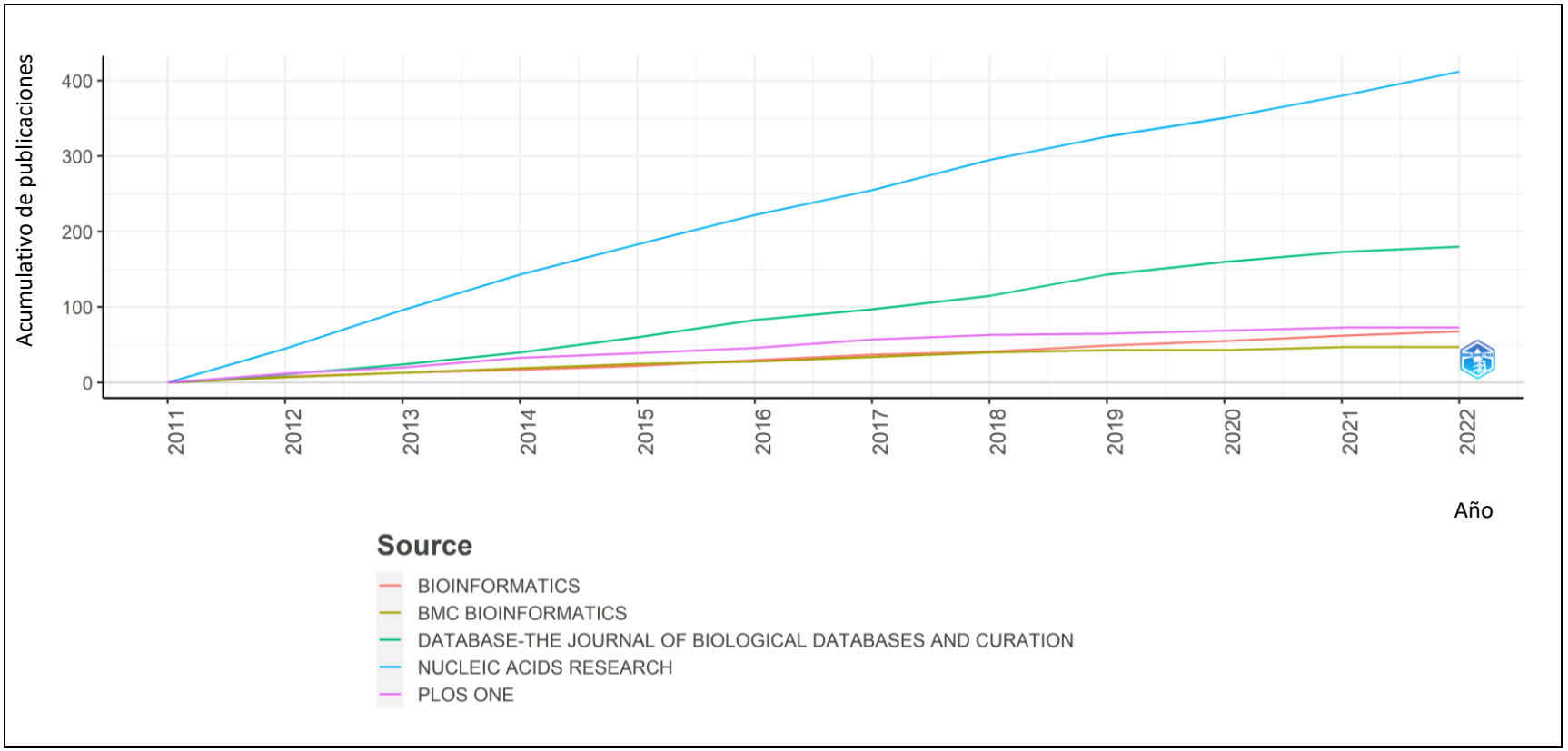

Gráfico 6. Crecimiento de las revistas acumulativo. Señala la productividad como el acumulativo a través de los años para las cinco revistas más relevantes.

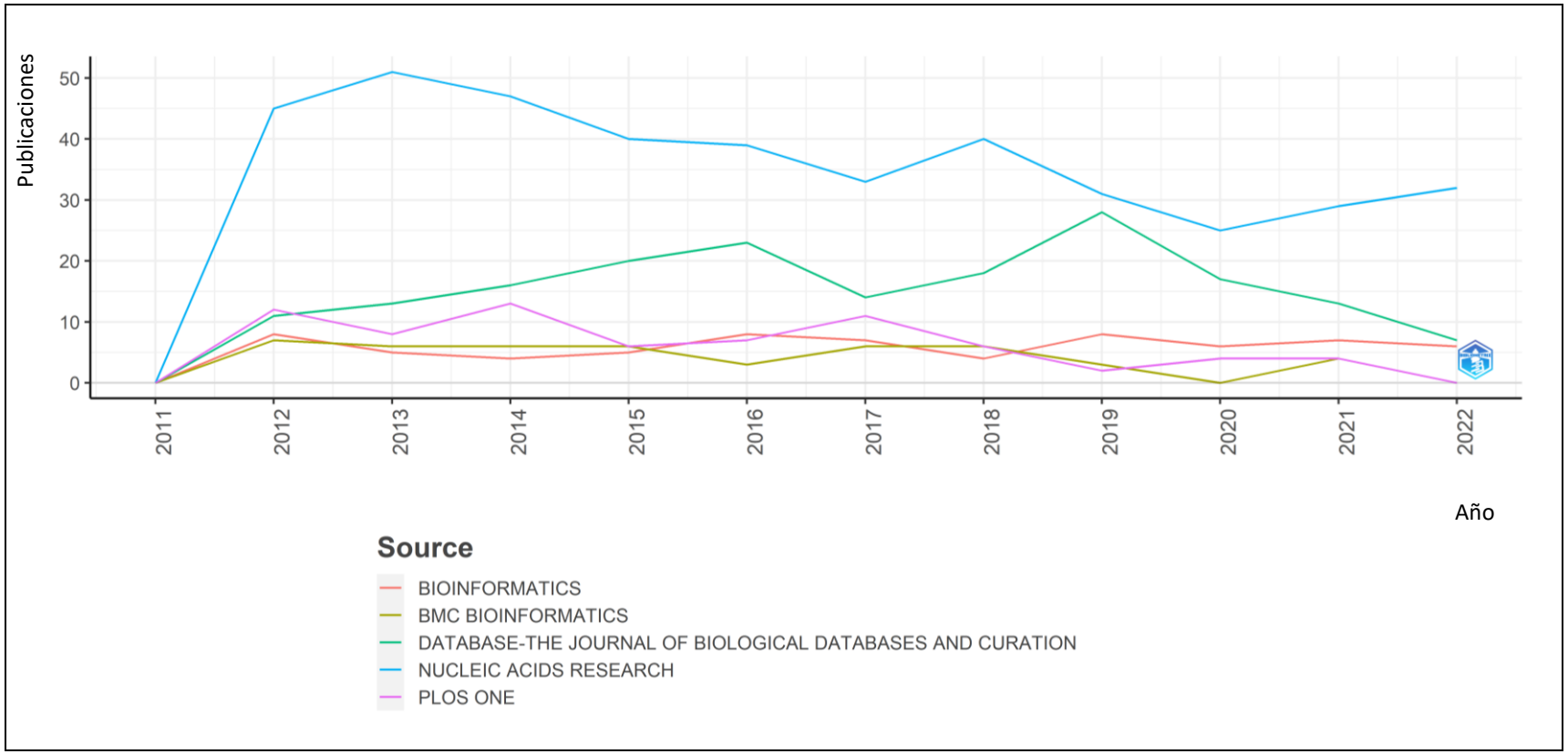

Gráfico 7. Crecimiento de las revistas por año. Señala la productividad año a año para las cinco revistas más relevantes.

Lo primero que salta a la vista en el Gráfico 6, sobre su crecimiento acumulativo, es que la revista del primer puesto en las más relevantes ya mencionada anteriormente —*Nucleic Acids Research*—, lo ha sido a través de todos los años medidos, y desde el principio ha mantenido su liderato a gran distancia del resto. Sin embargo, en el Gráfico 7, aunque también es la más productiva en cada uno de los años revisados, no se mantiene tan distante del resto siempre, hay años como el 2012 en que triplica o hasta sextuplica al resto, por períodos como el 2019 y 2020, se mantiene muy cerca de *Database: The Journal of Biological Databases and Curation*. Esta última se queda en el segundo puesto —según el acumulativo— de forma definitiva posterior al 2014, pues antes de esto los datos muestran una muy corta distancia entre las curvas para los puestos del segundo al quinto.

La distribución en la productividad año por año (Gráfico 7) permite ver cómo fue *PLOS ONE* la que ocupó el segundo puesto el 2012, y casi recuperándolo el 2014 y el 2017, pero por varios años disputando lugares con las revistas de los puestos cuarto y quinto.

Los lugares en productividad que alcanzan en diferentes años estas revistas no aseguran sus puestos de forma independiente en el acumulativo, solo las tendencias a lo largo del tiempo definen sus resultados a largo plazo. Así pues, como se mencionó al final del punto 8.4.3.2.1, *Bioinformatics* sigue de cerca la productividad de *PLOS ONE* —la que cierra hoy el Núcleo de Bradford—, pero ya el 2016 la superó, y también el 2019 y los años siguientes, por lo que no sería extraño que en unos años más si se mantiene igual o mejor en productividad podría quedarse ella con el tercer puesto de las revistas más relevantes.

7.4.3.4. *Revistas más citadas localmente*

Un aspecto diferente al tratado en el punto anterior, donde la cantidad de publicaciones era la variable dependiente, es el de las citas, pero no para medir al grupo de fuentes de esta colección, si no, en este caso, a las fuentes citadas por las publicaciones de esta colección. Esta información se obtiene al revisar las secciones de bibliografía al final de cada una de las publicaciones —Pudiendo ser varios miles—. Biblioshiny lo calcula de forma rápida el trabajar con el *dataset*, buscando específicamente en los metadatos que necesita, en este caso, en el campo referencias.

7.4.3.4.1. *Top 30 de las revistas más citadas localmente.*

El Gráfico 8, obtenido desde al menú *Sources/Local Cited Sources*, muestra las 30 revistas más citadas localmente. Los datos en él, el ranking y la distribución porcentual de las citaciones se muestran en la Tabla 19.

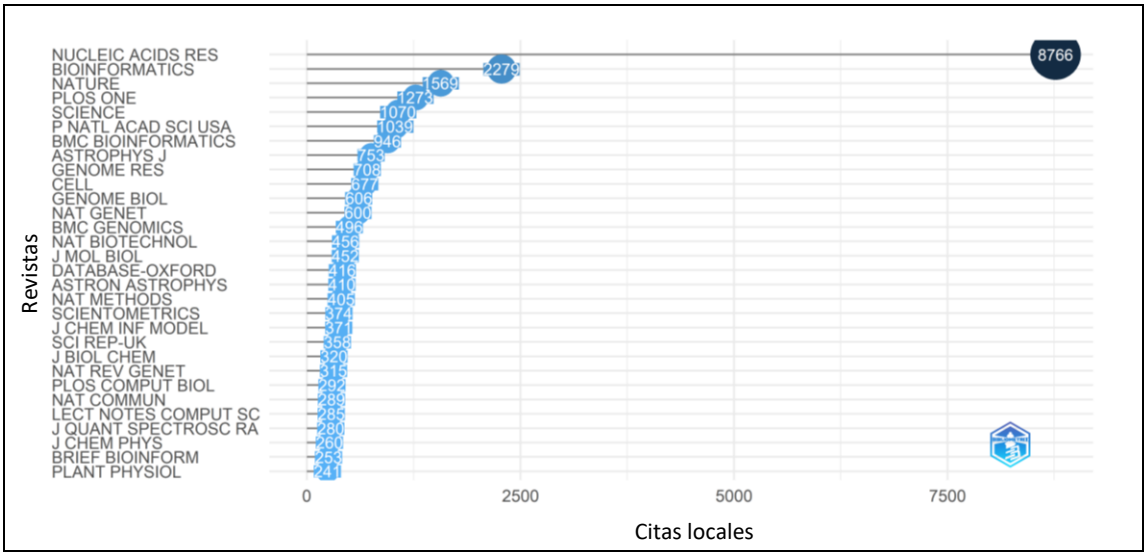

Gráfico 8. Revistas más citadas localmente. Las citas locales son aquellas hechas por los investigadores en las publicaciones analizadas en este *dataset*.

Tabla 19. Revistas más citadas localmente y distribución porcentual de las citas

Posición	*Revista*	**Citas Locales**	**%**
1	***Nucleic Acids Research***	**8766**	**10,89**
2	***Bioinformatics***	**2279**	**2,83**
3	***Nature****	**1569**	**1,95**
4	***PLOS ONE***	**1273**	**1,58**
5	***Science****	**1070**	**1,33**
6	*Proceedings of the National Academy of Sciences*	1039	1,29
7	*BMC Bioinformatics*	946	1,17
8	*Astrophysical Journal**	753	0,94
9	*Genome Research**	708	0,88
10	***Cell****	**677**	**0,84**
11	*Genome Biology**	606	0,75
12	*Nature Genetics**	600	0,75
13	*BMC Genomics*	496	0,62
14	*Nature Biotechnology**	456	0,57
15	*Journal of Molecular Biology*	452	0,56
16	***Database: The Journal of Biological Databases and Curation***	**416**	**0,52**
17	*Astronomy & Astrophysics*	410	0,51
18	*Nature Methods**	405	0,50
19	*Scientometrics*	374	0,46
20	*Journal of Chemical Information and Modeling*	371	0,46

Tabla 19. Revistas más citadas localmente y distribución porcentual de las citas

Posición	*Revista*	Citas Locales	%
21	*Scientific Reports*	358	0,44
22	*Journal of Biological Chemistry**	320	0,40
23	*Nature Reviews Genetics**	315	0,39
24	*PLOS Computational Biology*	292	0,36
25	*Nature Communications*	289	0,36
26	*Lecture Notes in Computer Science**	285	0,35
27	*Journal of Quantitative Spectroscopy and Radiative Transfer**	280	0,35
28	*Journal of Chemical Physics**	260	0,32
29	*Briefings in Bioinformatics*	253	0,31
30	*PLANT PHYSIOLOGY*	241	0,30
Total de citas locales en top 30		**26559**	**32,98%**
Total de citas locales en *dataset*		**80525**	
Total de fuentes citadas localmente		**15227**	

* Los nombres señalados fueron recuperados desde una fuente externa[15].

Al igual que en las revistas más relevantes (8.4.3.2), el 1° lugar de la lista de los más citados (ver Tabla 19) se lo lleva la revista *Nucleic Acids Research*, de la que provienen 8766 citas de un total de 80.525 (un 10,89%), siendo por lejos la revista más citada en las publicaciones de esta colección.

El 2° lugar en esta lista corresponde a *Bioinformatics*, revista también presente en la lista de las más relevantes, donde ocupó el cuarto lugar, y cuyo comportamiento especial se comentó al final del punto 8.4.3.2.1. El 4° lugar corresponde a *PLOS ONE*, revista tercera de entre las más relevantes, ocupando el puesto que cierra el Núcleo de Bradford.

Es interesante mencionar el intercambio de puestos que sufren estas dos revistas al comparar ambos gráficos, pues sin querer demostrar cuál de las dos variables influye más sobre la otra, "publicaciones en el tema" versus "citas locales", ambas figuran ocupando diferentes puestos de entre los primeros 4, sin llegar al primero, y cuando cambian no se alejan mucho de esta zona. Caso contrario es el de *Nucleic Acids Research*, la que además de mantener la primera posición, se mantiene muy por sobre los demás puestos en ambos gráficos.

[15] Estas revistas no presentan publicaciones en el *dataset*. Biblioshiny no entrega esta tabla con el metadato *Source*, correspondiente al nombre completo de la fuente, si no que con *Journal Abbreviation*, el que corresponde a la versión en mayúsculas y sin puntuación del metadato *Journal Abbreviation* ISO4, el que corresponde a la abreviación estándar oficialmente aceptada. Para presentar en esta tabla los nombres completos estos fueron recuperados desde el sitio https://academic-accelerator.com/Journal-Abbreviation/. El resto de los nombres en la columna fueron recuperados desde este mismo *dataset*.

Siguiendo el enfoque con respecto a los cambios en los primeros puestos, *Database: The Journal of Biological Databases and Curation*, que aparecía como segunda de entre las más relevantes, en esta lista de las más citadas localmente ocupa el 16° lugar. Ello indica que su alta especialización en productividad no lleva emparejado el mismo grado de interés por parte de los científicos —en este tema— al momento de citarla. Dicho lo anterior, cabe mencionar que este descenso desde el puesto 2 al 16, se da entre variables cuyas curvas describen un comportamiento logarítmico, por ende, en ambos rankings, ya estar en el top 30 —de un N de 469 para las más relevantes y 15227 para las más citadas localmente— representa una posición de relevancia en cuanto a producción o citación.

Así como se mencionó en los casos anteriores, que revistas en este listado de las más citadas localmente, también tienen presencia en las más relevantes, tenemos revistas que no aparecieron entre las más relevantes, o más aun, no presentan publicaciones en este *dataset*, por ende, podemos entender en tales casos que, aunque su especialización en el tema de las bases de datos web es poca o nula, sí son fuentes de referencia para las publicaciones de nuestro *dataset* y por ende para el tema. Ejemplo de ello son *Nature* (3° lugar), *Science* (5° lugar) y *Cell* (10° lugar). Las tres son revistas de renombre y alto reconocimiento internacional, figuran en esta lista como unas de las más citadas localmente, aun cuando no están enfocadas en el tema de esta investigación.

7.4.3.5. *Source Impact. Revistas con mayor impacto.*

A continuación, se mostrarán los "Top 30" según 4 medidas de impacto para las revistas de este *dataset*, calculadas en base a la cantidad de sus publicaciones, el tiempo transcurrido desde el inicio de estas y la cantidad de citas que hayan recibido, es por esto que las medidas de impacto aquí ocupadas son de carácter interno, y no reflejan necesariamente el impacto que puedan tener estas revistas en otras áreas y tampoco a nivel general[16]. Los Gráfico 9 y Gráfico 10 fueron obtenidos desde el menú *Sources/Sources Local Impact/Plot*, parámetro *Impact measure* en *H-Index* y *G-Index* respectivamente.

[16] En esta investigación se evalúa un *dataset* bajo diferentes parámetros de medida, todos ellos basados en los metadatos correspondientes a los registros bibliográficos de las publicaciones detectadas con los parámetros de búsqueda especificados en la metodología (punto 7.1). Dichos parámetros caracterizan al *dataset*, y con ello caracterizan también a los resultados de todas las mediciones hechas, mostradas y discutidas en esta tesis. Ello aplica también, por supuesto, a las medidas de impacto mostradas y discutidas en este punto.

7.4.3.5.1. *Revistas con mayor índice H.*

Tomamos el índice H como medida de impacto, índice que da cuenta de la cantidad H de publicaciones con por lo menos una cantidad H de citas (ver punto 3.2.2.2.1), mostrando ello en un solo dato una relación cantidad-calidad.

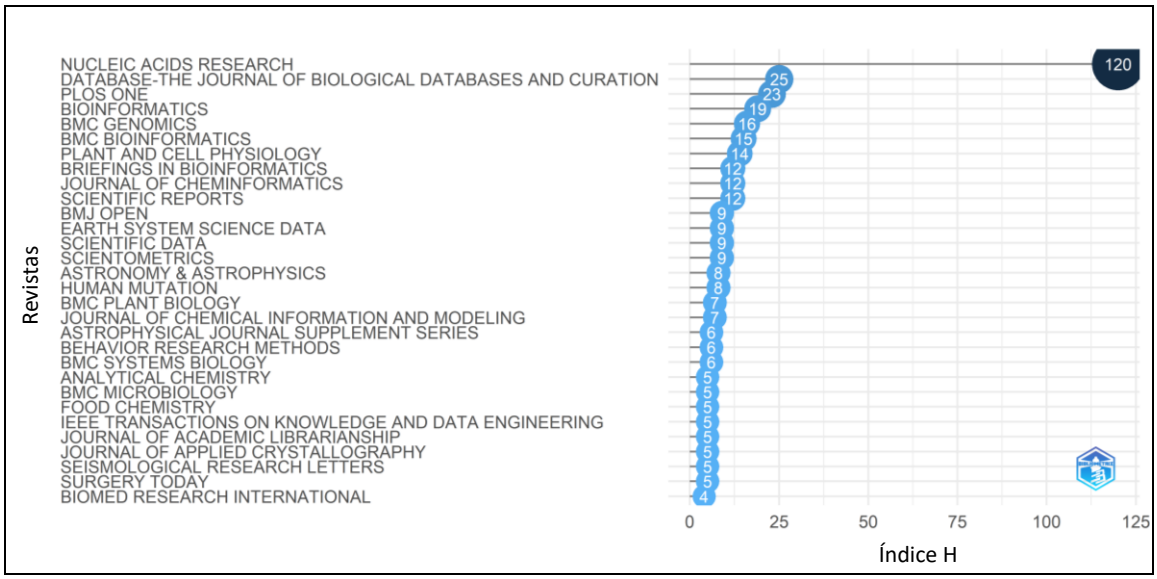

Gráfico 9. Top 30 de las revistas según índice H interno

Como se observa en el Gráfico 9, las 4 revistas con mayor impacto según su índice H son:

1° *Nucleic Acids Research* (120)

2° *Database: The Journal of Biological Databases and Curation* (25)

3° *PLOS ONE* (23)

4° *Bioinformatics* (19)

Ellas coinciden además al ser las mismas 4 con mayor productividad (Gráfico 4), además manteniendo sus mismos puestos según aquella medida. El índice H de *Nucleic Acids Research* es de 120, muy por lejos del resto, siguiéndole *Database: The Journal of Biological Databases and Curation* con 25. Desde el índice H del puesto 2 en adelante, el descenso de los valores desciende con bajo ritmo.

Para enriquecer más el análisis de impacto de las revistas, se incluyen a continuación 4 tablas con la lista de los "Top 30" según el factor de impacto señalado, mostrando además el resto de las medidas de impacto. Estas tablas de números **Tabla 20**, **Tabla 21**, **Tabla 22** y **Tabla 24** son

en realidad variaciones de una única tabla obtenida desde el menú *Sources/Source Local Impact/Table*, ordenando los registros según **H**, **G**, ***m*** y **TC** respectivamente. En las Tabla 20, Tabla **21** y Tabla **22**, las posiciones compartidas por igual valor de H, G y *m* respectivamente, fueron ordenadas de forma secundaria según TC[17].

Tabla 20. '*Source Impact*' Top 30 según índice H.

Posición	Revista	Índice H	Índice G	Cociente *m*	TC*	NP**	Año inicial
1	***Nucleic Acids Research***	**120**	**304**	**10,909**	**93413**	**406**	**2012**
2	*Database: The Journal of Biological Databases and Curation*	**25**	43	2,273	2619	153	2012
3	*PLOS ONE*	**23**	44	2,091	2132	69	2012
4	*Bioinformatics*	**19**	47	1,727	2233	56	2012
5	*BMC Genomics*	**16**	28	1,333	840	34	2011
6	*BMC Bioinformatics*	**15**	24	1,364	693	42	2012
7	*Plant and Cell Physiology*	**14**	21	1,273	1102	21	2012
8	Scientific Reports	**12**	26	1,2	743	26	2013
	Briefings in Bioinformatics	**12**	24	1,2	737	24	2013
	Journal of Cheminformatics	**12**	19	1,2	648	19	2013
9	Scientific Data	**9**	20	1,125	547	20	2015
	Scientometrics	**9**	21	0,818	483	23	2012
	Earth System Science Data	**9**	13	1	284	13	2014
	BMJ Open	**9**	13	0,818	174	15	2012
10	*Human Mutation*	**8**	12	0,727	931	12	2012
	Astronomy & Astrophysics	**8**	8	0,8	429	8	2013
11	*BMC Plant Biology*	**7**	8	0,636	248	8	2012
	Journal of Chemical Information and modeling	**7**	13	0,583	188	17	2011
12	*Behavior Research Methods*	**6**	8	0,75	216	8	2015
	Astrophysical Journal Supplement Series	**6**	6	0,667	211	6	2014
	BMC Systems Biology	**6**	8	0,545	143	8	2012
13	*Analytical Chemistry*	**5**	5	0,455	207	5	2012
	Journal of Applied Crystallography	**5**	5	0,625	205	5	2015
	Surgery Today	**5**	5	0,625	148	5	2015
	BMC Microbiology	**5**	5	0,455	135	5	2012
	Seismological Research Letters	**5**	8	0,5	125	8	2013
	Food Chemistry	**5**	6	0,5	100	6	2013
	Journal of Academic Librarianship	**5**	5	0,5	64	5	2013
	IEEE Transactions on Knowledge and Data Engineering	**5**	7	0,455	55	7	2012
14***	*VLDB Journal****	**4**	5	0,364	180	5	2012

* Total de citas

** Cantidad de publicaciones, solo incluye aquellas con por lo menos una cita.

*** La posición 14 es compartida por 14 revistas en total.

[17] Este valor del total de citas (TC) que se menciona en las 5 secciones de este punto 8.4.3.4, no debe confundirse con las citas locales (LC), tratadas en el punto 8.4.3.3.

El orden de las revistas desde el 5° puesto en adelante es diferente al presentado en el (Gráfico 4). Ello corresponde a que, en este caso, el índice H representa una medida basada en dos parámetros diferentes: cantidad de publicaciones y cantidad de citas de las mismas. Por ejemplo, justo desde ese nivel en la lista, *BMC Genomics* (5°) y *BMC Bioinformatics* (6°), intercambian los puestos que mostraron en la lista de las más relevantes, pues, aunque *BMC Bioinformatics* cuenta con 42 publicaciones en el tema, según al cálculo cantidad-calidad que busca con su fórmula el índice H, *BMC Genomics* habría generado un mayor impacto[18].

Para referirse a la parte baja de esta lista, los lugares en realidad están bien definidos por H solo hasta la 13° posición, ya que al ser H un valor discreto, conforme se desciende en la lista, y decrece el valor de H, este se comienza a repetir, compartiendo posición en el ranking un número cada vez mayor de revistas. La posición 8 la comparten 3 revistas, la posición 13 la comparten 8 revistas, y hasta aquí la lista va en 29. Si bien esta tabla corresponde a "las 30 más", la posición 14, lugar que ocupa la última revista mostrada en esta tabla, es compartida en realidad por otras 13 revistas más.

Estas 30 revistas con mayor impacto según su índice H, se describen por índices $H \geq 4$ y $G \geq 5$, cociente $m \geq 0{,}4$, total de citas ≥ 55, y por lo menos 5 publicaciones en el tema.

7.4.3.5.2. *La revista top en índices H, G, cociente m y TC.*

La única revista en mantener su posición en estas 4 medidas de impacto es la primera en la lista, *Nucleic Acids Research*, la que además de liderar en índice H, lo hace también en índice G, cociente *m*, y TC —además de sus lideratos en relevancia y citación interna comentados anteriormente en los puntos 8.4.3.2.1 y 8.4.3.3.1—.

El índice H de 120 indica que la revista ha liberado 120 publicaciones en el tema de las bases de datos web que lograron ya a la fecha de corte de esta investigación por lo menos 120 citas cada una, correspondiendo estas a citas globales, es decir, sin distinción de tema.

Este índice H de 120 no sería tal sin que hayan transcurrido todo el período de tiempo evaluado, sin embargo, tras verlo de forma anualizada con su cociente *m* de 10,909, tomando en

[18] Esto no debe interpretarse como una mayor influencia de las 34 publicaciones que ha producido, pues solo indica que 16 de ellas tienen por lo menos 16 citas, sin influir en el valor total de H las 34 publicaciones, tampoco su valor total de citas.

cuenta el período evaluado en esta investigación, desde ese año como inicio registral de sus publicaciones, ha mantenido un índice de impacto equivalente a un índice H ≈ 11 obtenido en solo un año.

El índice H de 120 es muy alto en comparación al resto, sin embargo, es una medida de los mínimos, es decir, son 120 las publicaciones que tienen por lo menos 120 citas, pero no dice de hasta cuánto pueden ser estas. El índice G de 304 nos da a entender que más allá de los mínimos, si se toma en cuenta también el aporte del total de citas de sus publicaciones más citadas, mayor aun es el impacto.

El total de las citas de 93413 corresponde a las citas que han recibido las publicaciones de esta revista —presentes en este *dataset*— desde cualquier área (citas globales). Si tomamos en cuenta el conteo de citas locales, es decir, las veces que alguna de las publicaciones de este *dataset* ha citado a la revista, 8766 citas locales (Tabla 19), son más de 10 veces las citas en todas las áreas que las citaciones que esta colección[19] hace a la revista. Esto indica que, por lo menos en cuanto a las citas, esta revista no solo tiene un gran impacto en el tema central de esta investigación, las bases de datos web, si no que el conjunto de publicaciones en este tema aquí analizadas también tiene un gran impacto en el resto de las áreas de investigación.

7.4.3.5.3. *Revistas con mayor índice G.*

Los valores de G, indican que si además de la relación cantidad-calidad del cálculo se toman en cuenta las cantidades totales de citas de las publicaciones más citadas. Según ello se grafican las 30 revistas con mayor impacto según el índice G:

[19] El conjunto de publicaciones en este *dataset*.

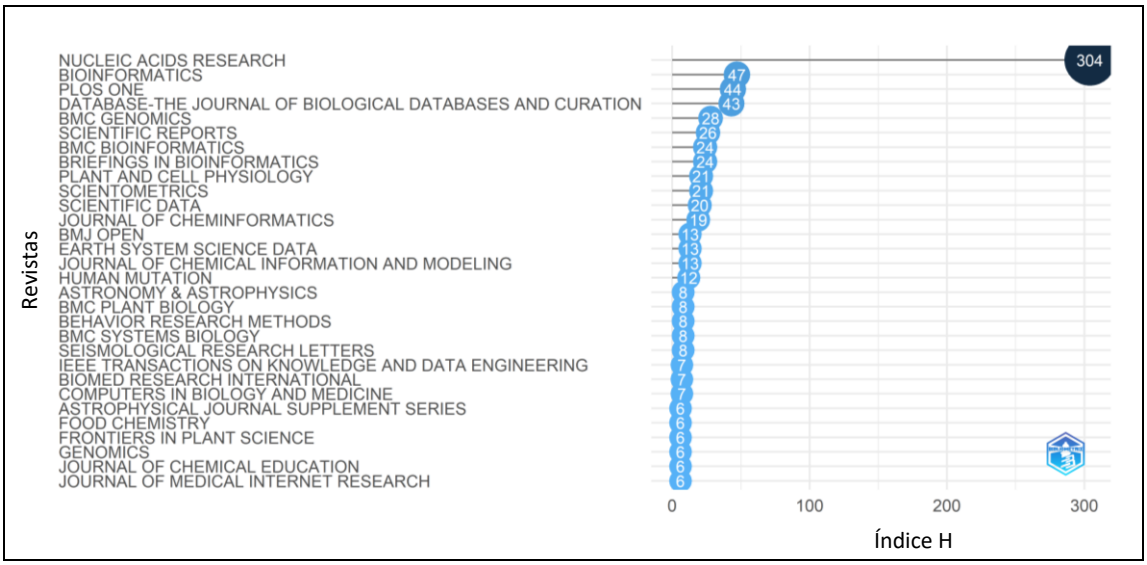

Gráfico 10. Top 30 de las revistas según índice G interno

Los cuatro primeros puestos están conformados por las mismas cuatro revistas top en el índice H, salvo que, en este caso, *Database: The Journal of Biological Databases and Curation* ya no ocupa el segundo puesto si no que el cuarto, y a su vez, *Bioinformatics* sube del cuarto puesto al segundo.

1° *Nucleic Acids Research* (304)

2° *Bioinformatics* (47)

3° *PLOS ONE* (44)

4° *Database: The Journal of Biological Databases and Curation* (43)

Al observar el Gráfico 10, resalta que las revistas de los puestos segundo tercero y cuarto, figuran con casi los mismos valores de impacto, diferenciándose bastante de *BMC Genomics* en el puesto quinto y las siguientes, a diferencia del descenso más bien continuo visto anteriormente para los valores de H en el Gráfico 9.

Este índice al igual que H también se expresa solo como valores discretos, por lo que también sucede que muchas revistas se igualan en posición, aunque en este top 30 se aprecian 15 posiciones diferentes según G, más que en el top 30 según H.

Tabla 21. '*Source Impact*' Top 30 según índice G.

Posición	Revista	Índice H	Índice G	Cociente *m*	TC*	NP**	Año inicial
1	***Nucleic Acids Research***	**120**	**304**	**10,909**	**93413**	**406**	**2012**
2	*Bioinformatics*	19	**47**	1,727	2233	56	2012
3	*PLOS ONE*	23	**44**	2,091	2132	69	2012
4	*Database: The Journal of Biological Databases and Curation*	25	**43**	2,273	2619	153	2012
5	*BMC Genomics*	16	**28**	1,333	840	34	2011
6	*Scientific Reports*	12	**26**	1,200	743	26	2013
7	*Briefings in Bioinformatics*	12	**24**	1,200	737	24	2013
	BMC Bioinformatics	15	**24**	1,364	693	42	2012
8	*Plant and Cell Physiology*	14	**21**	1,273	1102	21	2012
	Scientometrics	9	**21**	0,818	483	23	2012
9	*Scientific Data*	9	**20**	1,125	547	20	2015
10	*Journal of Cheminformatics*	12	**19**	1,200	648	19	2013
11	*Earth System Science Data*	9	**13**	1,000	284	13	2014
	Journal of Chemical Information and Modeling	7	**13**	0,583	188	17	2011
	BMJ Open	9	**13**	0,818	174	15	2012
12	*Human Mutation*	8	**12**	0,727	931	12	2012
13	*Astronomy & Astrophysics*	8	**8**	0,800	429	8	2013
	BMC Plant Biology	7	**8**	0,636	248	8	2012
	Behavior Research Methods	6	**8**	0,750	216	8	2015
	BMC Systems Biology	6	**8**	0,545	143	8	2012
	Seismological Research Letters	5	**8**	0,500	125	8	2013
14	*Biomed Research International*	4	**7**	0,400	62	7	2013
	Computers in Biology and Medicine	4	**7**	0,444	60	7	2014
	IEEE Transactions on Knowledge and Data Engineering	5	**7**	0,455	55	7	2012
15	*Astrophysical Journal Supplement Series*	6	**6**	0,667	211	6	2014
	Journal of Medical Internet Research	3	**6**	0,429	171	6	2016
	Genomics	4	**6**	0,364	116	6	2012
	Food Chemistry	5	**6**	0,500	100	6	2013
	Frontiers in Plant Science	4	**6**	0,364	97	6	2012
	Journal of Chemical Education	4	**6**	0,364	43	6	2012

* Total de citas

** Cantidad de publicaciones, solo incluye aquellas con por lo menos una cita.

Estas 30 revistas con mayor impacto según su índice G, se describen por índices $H \geq 3$ y $G \geq 6$, cociente $m \geq 0{,}364$, total de citas ≥ 43, y por lo menos 6 publicaciones en el tema.

7.4.3.5.4. *Revistas con mayor cociente m.*

Según la Tabla 22, las 4 revistas con mayor impacto según su cociente *m* son:

1° *Nucleic Acids Research* (10,909)

2° *Database: The Journal of Biological Databases and Curation* (2,273)

3° *PLOS ONE* (2,091)

4° *Bioinformatics* (1,727)

Las primeras 11 revistas en este listado visible en la Tabla 22 —desde *Nucleic Acids Research* hasta *Scientific Data*—cuentan con índices H, G, cociente *m*, TC, NP y hasta el año inicial con los mismos valores que los ya mostrados para las primeras 11 revistas en la Tabla 22, donde el orden estaba regido por H, y es que son las 11 revistas que lideran en ese índice de impacto y en este. Todas las revistas de este primer tramo muestran un cociente $m > 1$.

Tabla 22. '*Source Impact*' Top 30 según cociente *m.*

Posición	Revista	Índice H	Índice G	Cociente *m*	TC*	NP**	Año inicial
1	*Nucleic Acids Research*	120	304	**10,909**	93413	406	2012
2	*Database: The Journal of Biological Databases and Curation*	25	43	**2,273**	2619	153	2012
3	*PLOS ONE*	23	44	**2,091**	2132	69	2012
4	*Bioinformatics*	19	47	**1,727**	2233	56	2012
5	*BMC Bioinformatics*	15	24	**1,364**	693	42	2012
6	*BMC Genomics*	16	28	**1,333**	840	34	2011
7	*Plant and Cell Physiology*	14	21	**1,273**	1102	21	2012
8	*Scientific Reports*	12	26	**1,200**	743	26	2013
	Briefings in Bioinformatics	12	24	**1,200**	737	24	2013
	Journal of Cheminformatics	12	19	**1,200**	648	19	2013
9	*Scientific Data*	9	20	**1,125**	547	20	2015
10	*Earth System Science Data*	9	13	**1,000**	284	13	2014
	International Journal of Environmental Research and Public Health	3	3	**1,000**	32	3	2020
	Journal of Information Science	3	3	**1,000**	20	3	2020
	European Journal of Pediatrics	1	1	**1,000**	2	1	2022
	Heritage Science	1	1	**1,000**	2	1	2022
	Nature Human Behaviour	1	1	**1,000**	2	1	2022
11	*Scientometrics*	9	21	**0,818**	483	23	2012
	BMJ Open	9	13	**0,818**	174	15	2012
12	*Astronomy & Astrophysics*	8	8	**0,800**	429	8	2013
13	*Behavior Research Methods*	6	8	**0,750**	216	8	2015
	Journal of Molecular Biology	3	4	**0,750**	36	4	2019
14	*Human Mutation*	8	12	**0,727**	931	12	2012
15	Astrophysical Journal Supplement Series	6	6	**0,667**	211	6	2014
	Nature Communications	2	2	**0,667**	50	2	2020
	Life Sciences	2	2	**0,667**	7	2	2020
	Medical Science Monitor	2	2	**0,667**	4	2	2020
16	*BMC Plant Biology*	7	8	**0,636**	248	8	2012
17	*Journal of Applied Crystallography*	5	5	**0,625**	205	5	2015
18	*Surgery Today*	5	5	**0,625**	148	5	2015

* Total de citas

** Cantidad de publicaciones, solo incluye aquellas con por lo menos una cita.

Todas las revistas del segundo tramo muestran un cociente $m = 1$. En este tramo, solo *Earth System Science Data* —la primera revista de este tramo— es también comparable en posición y el resto de sus indicadores a las listas anteriores. El resto de las revistas en este tramo,

difiere por mucho de los listados top 30 ya revisados para H y G. Las dos siguientes revistas comenzaron a publicar el 2020, y las otras tres, recién este 2022. Estas últimas 5 revistas, se verían totalmente invisibilizadas por los listados top 30 según H y G. Aun cuando los valores para sus diferentes indicadores son mínimos, ellas muestran un mayor desempeño según la relación de impacto cantidad-calidad-temporalidad medida por el indicador de impacto cociente *m* que el resto de las revistas en el tercer tramo de este listado.

La característica de este indicador de no expresarse necesariamente como un valor discreto, entrega una mayor definición al momento de separar las revistas según su posición en el listado, sin embargo, cuando el año registral de la revista no es mayor a uno, varias revistas coinciden en su valor de *m*. Si se quiere profundizar más el análisis de impacto en tramos del listado que muestren igual valor de *m*, pueden en tales casos revisarse sus demás indicadores.

El comienzo del tercer tramo de este listado, muestra a revistas como *Scientometrics*, *BMJ Open*, y varias más también presentes en los top 30 para H y G. Dos lugares más abajo, se encuentra *Behavior Research Methods*, con cociente *m* = 0,750 al igual que *Journal of Molecular Biology*, lo que indica que ambas revistas muestran igual impacto aun cuando la primera tiene un H del doble que la segunda, esto ya que la segunda lleva la mitad del tiempo publicando en el tema.

Este listado según el cociente *m*, muestra a revistas que, aunque lleven poco tiempo publicando en el tema, con el paso de los años podrían llegar a tener el mismo impacto que otras que las igualan en *m*, si mantuviesen la performance medida hasta ahora a través de los siguientes años.

Tabla 23. Cantidad de revistas por trienio en rankings top 30 según cada indicador

Edad de la revista	Índice H	Índice G	Cociente *m*	TC
2013 o antes	24	24	15	18
2014 - 2016	6	6	6	8
2017 - 2019	0	0	1	4
2020 - 2022	0	0	8	0

La edad de la revista —interna por supuesto en este caso—, implica una distribución diferente de los rankings de evaluación de impacto según en qué variable se base cada uno de ellos. Los índices H y G, privilegian las publicaciones del 2013 o antes, un 80%, y el otro 20%

para el período 2014 – 2016, dejando ambos rankings totalmente fuera a revistas que hubiesen comenzado a publicar en este tema hace menos tiempo. Es el cociente m el que incluye a revistas que resaltan en impacto si es considerado el año en que iniciaron sus publicaciones en el tema, mostrando en este caso una alta cantidad de revistas que destacan cuyo año de inicio registral es del 2020 o posterior (ver en Tabla 23).

7.4.3.5.5. *Revistas con mayor TC.*

Los tres listados top 30 mostrados en los puntos anteriores, fueron hechos en base a índices de impacto con relaciones cantidad-calidad, siendo el índice G y el cociente *m* derivaciones directas del índice H, el primero mostrado. En el listado a mostrar a continuación, se tomó en cuenta solo el factor citaciones, la cantidad de publicaciones no está tomada en cuenta a diferencia del cálculo de los indicadores tomados en cuenta en los tres anteriores listados.

Tabla 24. '*Source Impact'* Top 30 según total de citas (TC).

Posición	Revista	Índice H	Índice G	Cociente *m*	TC*	NP**	Año inicial
1	***Nucleic Acids Research***	**120**	**304**	**10,909**	**93413**	**406**	**2012**
2	*Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*	1	1	0,143	**5360**	1	2016
3	*Database: The Journal of Biological Databases and Curation*	25	43	2,273	**2619**	153	2012
4	*Journal of Neuroscience*	2	2	0,222	**2618**	2	2014
5	*IEEE Transactions on Affective Computing*	3	3	0,273	**2383**	3	2012
6	*Bioinformatics*	19	47	1,727	**2233**	56	2012
7	*PLOS ONE*	23	44	2,091	**2132**	69	2012
8	*Plant and Cell Physiology*	14	21	1,273	**1102**	21	2012
9	*RNA*	2	3	0,222	**1008**	3	2014
10	*Journal of Atmospheric and Oceanic Technology*	1	1	0,091	**944**	1	2012
11	*Human Mutation*	8	12	0,727	**931**	12	2012
12	*BMC Genomics*	16	28	1,333	**840**	34	2011
13	*Scientific Reports*	12	26	1,2	**743**	26	2013
14	*Briefings in Bioinformatics*	12	24	1,2	**737**	24	2013
15	*BMC Bioinformatics*	15	24	1,364	**693**	42	2012
16	*Physical Chemistry Chemical Physics*	1	1	0,167	**688**	1	2017
17	*Journal of Cheminformatics*	12	19	1,2	**648**	19	2013
18	*Human and Ecological Risk Assessment*	1	1	0,143	**563**	1	2016
19	*Scientific Data*	9	20	1,125	**547**	20	2015
20	*Gut*	3	3	0,5	**540**	3	2017
21	*Fungal Diversity*	1	1	0,125	**522**	1	2015
22	*Scientometrics*	9	21	0,818	**483**	23	2012
23	*Publications of the Astronomical Society of The Pacific*	3	3	0,333	**469**	3	2014

Tabla 24. '*Source Impact'* Top 30 según total de citas (TC).

Posición	Revista	Índice H	Índice G	Cociente *m*	TC*	NP**	Año inicial
24	*Molecular Biology and Evolution*	1	1	0,1	**456**	1	2013
25	*Astronomy & Astrophysics*	8	8	0,8	**429**	8	2013
26	*Systematic Reviews*	1	1	0,167	**378**	1	2017
27	*Journal of Biotechnology*	2	2	0,333	**336**	2	2017
28	*BMC Medical Research Methodology*	3	4	0,3	**323**	4	2013
29	*Solar Energy*	1	1	0,091	**316**	1	2012
30	*Annals of Surgery*	1	1	0,111	**314**	1	2014

* Total de citas

** Cantidad de publicaciones, solo incluye aquellas con por lo menos una cita.

Como se observa en la Tabla 24, las 4 revistas con mayor impacto según su total de citas (TC) son:

1° *Nucleic Acids Research* (93.413)

2° *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* (5.360)

3° *Database: The Journal of Biological Databases and Curation* (2.619)

4° *Journal of Neuroscience* (2.618)

En este caso, al tomar en cuenta solo el valor de TC para cada revista, con la excepción ya mencionada en el punto 8.4.3.4.2 para la única revista que se mantiene en su posición y liderato en todas las mediciones de impacto consideradas, *Nucleic Acids Research*, este listado, muestra grandes diferencias en el orden de la tabla.

Estos reordenamientos se deben a la asombrosa cantidad de citas de revistas considerando su bajísima cantidad de publicaciones aquí contabilizadas. Ejemplo de ello es la revista que ocupa el 2° puesto de este ranking, *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*, cuya altísima cantidad de citas (5360) corresponde a un solo artículo[20]. Casos similares son las revistas *Journal of Neuroscience* (2.618), *IEEE Transactions on Affective Computing* (2383) y *RNA* (1008), con 2, 3 y 3 publicaciones respectivamente.

[20] Esta publicación corresponde a la presentación de la base de datos de moléculas "*The Cambridge Structural Database (CSD)* Groom, C. R., Bruno, I. J., Lightfoot, M. P., & Ward, S. C. (2016). The Cambridge Structural Database. *Acta Crystallographica Section B, 72*(2), 171-179. https://doi.org/10.1107/S2052520616003954 . No se profundiza en ella por el enfoque del análisis bibliométrico en esta investigación, puesto en la epistemometría de las revistas científicas del área de las bases de datos. Solo se hace referencia al artículo de forma excepcional por el alto impacto demostrado con su 2° puesto como la revista más citada siendo de publicación única, sin embargo, no es el artículo más citado, lo preceden dos y prosiguen otros tres artículos de la revista *Nucleic Acids Research*.

8. Conclusiones

El método de deduplicación de registros bibliográficos realizado con EndNote X9, demuestra ser simple, rápido y útil al estar integrado en un gestor de referencias asociado a la base datos WoS, la misma utilizada para la obtención de los datos, sin embargo, en cuanto a la cantidad de duplicados encontrados, es menos eficaz que el método por Excel, el que si bien, es mucho menos automatizado que el anterior, permite un análisis comparativo más acucioso de los campos bibliográficos claves en este proceso, logrando identificar —además de los duplicados que ya encuentra EndNote X9— registros bibliográficos en duplicado que no encuentra el primero.

El barrido por impacto realizado al *dataset* original, en base al índice F_H de las revistas científicas, logra reducir en un porcentaje importante la cantidad de publicaciones del *dataset* original, sin perder en este proceso registros de publicaciones con alto impacto —en el área de investigación de las bases de datos web— según su índice H, cantidad de citas, e incluso considerando en ambos aspectos el tiempo transcurrido para la medida de ambos, generando un insumo útil para análisis bibliométrico, sobre todo cuando este pone el foco en describir los aspectos destacables de la producción científica en una temática particular, ya que su efecto se traduce en reducir los registros de publicaciones que no resalten en ninguno de los aspectos señalados como indicadores de impacto.

Este *dataset* producto del barrido por impacto según el índice F_H, representa un insumo de vital importancia para la continuación de la investigación bajo el marco de una revisión sistemática. Tanto el *dataset* completo como las tablas resultantes del análisis bibliométrico se encuentran disponibles en formato de hoja de cálculo (.xlsx). Para posteriores análisis en Biblioshiny, se cuenta con el *dataset* en un archivo apto para ser cargado y analizado directamente por dicha plataforma.

El análisis bibliométrico realizado a nivel de fuentes, permite identificar a las revistas más influyentes en el campo de las bases de datos. Las 4 revistas más influyentes según su índice H son también las más productivas, pues solo de la primera proviene más del 21% de las publicaciones del tema, y estas cuatro en su conjunto producen casi el 38% de todas las publicaciones, además, un caso particular en cuanto a producción es el de la editorial de la

Universidad de Oxford, que a través de las revistas bajo su edición concentra un porcentaje similar al recién descrito, pero no de tan alto impacto. *Nucleic Acids Research* es por lejos la revista más productiva, la más citada —global y localmente, siendo más citada por los autores en este *dataset* incluso por encima de otras revistas de renombre como *Nature* o *Science*—, además mantiene el liderato no solo en el índice H, sino también en sus derivados índice G y cociente *m*, mostrando gran performance en todos los indicadores medidos en este estudio. Las revistas que encabezan el índice H también lo hacen según el cociente *m*, pero este último indicador muestra una sobre representación para las revistas del último trienio. Una medición de impacto según el índice G muestra modifica la tabla subiendo y bajando algunos puestos a ciertas revistas según sus publicaciones más citadas, pero los primeros lugares son bastante parecidos.

Al medir a las revistas solo por el total de citas acumulado por sus publicaciones (TC), es posible encontrar en los primeros lugares a revistas que no muestran trayectoria en el campo de las bases de datos científicas, pues registran solo una o muy pocas publicaciones en el tema, y aun así muestran una altísima cantidad de citaciones. Ejemplo de ello es la revista *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*, cuya única publicación en el *dataset* registra 5.360 citas, ocupando el segundo lugar en las revistas más citadas, solo después de *Nucleic Acids Research*.

El análisis bibliométrico se limita más a lo cuantitativo que a lo cualitativo, sin embargo, aporta conocimiento clave en la construcción del estado del arte del tema de estudio. Preguntas como ¿cuáles son los autores más influyentes en el campo de las bases de datos científicas en línea? ¿cuáles son sus redes de trabajo? ¿qué universidades son las principales contribuyentes al desarrollo de este campo? ¿en qué se enfocan las principales bases de datos científicas en línea? ¿son todas de acceso abierto? ¿cuáles de estas son más idóneas para su utilización resolviendo problemáticas reales en entornos de aprendizaje STEM? son abordables al continuar un análisis bibliométrico a niveles de autor, y sobre todo de documentos, pero esto implica una extensión en tiempo mucho mayor al disponible. El enfoque bibliométrico y epistemométrico del análisis realizado, sienta las bases para la prosecución de la investigación, que se proyecta desde una metodología de revisión sistemática, para abordar preguntas de investigación de carácter más cualitativo que escapan de lo bibliométrico.

9. Referencias

Arencibia-Jorge, R., & de Moya-Anegon, F. (2008). La evaluación de la investigación científica: una aproximación teórica desde la Cienciometría. *Acimed: Revista cubana de los profesionales de la información y la comunicación en salud, 17*, 1-27. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1024-94352008000400004

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics, 11*(4), 959-975. https://doi.org/10.1016/j.joi.2017.08.007

Berrington, J. (2017). Databases. *Anaesthesia & Intensive Care Medicine, 18*(3), 155-157. https://doi.org/10.1016/j.mpaic.2016.11.016

Bornmann, L., Mutz, R., & Daniel, H.-D. (2008). Are there better indices for evaluation purposes than theh index? A comparison of nine different variants of theh index using data from biomedicine. *Journal of the American Society for Information Science and Technology, 59*(5), 830-837. https://doi.org/10.1002/asi.20806

Bornmann, L., Mutz, R., Hug, S. E., & Daniel, H.-D. (2011). A multilevel meta-analysis of studies reporting correlations between the h index and 37 different h index variants. *Journal of Informetrics, 5*(3), 346-359. https://doi.org/10.1016/j.joi.2011.01.006

Bradford, S. C. (1985). Sources of information on specific subjects 1934 [Re-edition]. *Journal of Information Science, 10*(4), 176-180. https://doi.org/10.1177/016555158501000407

Braun, T., Glänzel, W., & Schubert, A. (2006). A Hirsch-type index for journals. *Scientometrics, 69*(1), 169-173. https://doi.org/10.1007/s11192-006-0147-4

Brookes, B. C. (1969). Bradford's Law and the Bibliography of Science. *Nature, 224*(5223), 953-956. https://doi.org/10.1038/224953a0

Camps, D. (2007). Estudio bibliométrico general de colaboración y consumo de la información en artículos originales de la revista Universitas Médica, período 2002 a 2006. *Universitas Médica, 48*(4), 358-365. https://www.redalyc.org/articulo.oa?id=231018670002

Clarivate Analytics. (2022, 10-06). *Web of Science: Definition and Use of Wildcards*. https://support.clarivate.com/ScientificandAcademicResearch/s/article/Web-of-Science-Definition-and-Use-of-Wildcards?language=en_US

Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM, 13*(6), 377-387. https://doi.org/10.1145/362384.362685

Cornella, A. (2000). *Cómo sobrevivir a la infoxicación*. https://web.archive.org/web/20190429043743id_/http://www.infonomia.com/img/pdf/sobrevivir_infoxicacion.pdf

Date, C. J. (2001). *Introducción a los sistemas de bases de datos*. Pearson Educación.

Díaz, I., Cortey, M., Olvera, À., & Segalés, J. (2016). Use of H-Index and Other Bibliometric Indicators to Evaluate Research Productivity Outcome on Swine Diseases. *PLOS ONE, 11*(3), e0149690. https://doi.org/10.1371/journal.pone.0149690

DOI Chile. (s.f.). *DOI – Digital Object Identifier para Latinoamérica*. Retrieved 31-06-2022 from https://doichile.cl/

Egghe, L. (2006a). An improvement of the h-index: The g-index. *ISSI Newsletter, 2*(1), 8-9. http://pds4.egloos.com/pds/200703/08/11/g_index.pdf

Egghe, L. (2006b). Theory and practise of the g-index. *Scientometrics, 69*(1), 131-152. https://doi.org/10.1007/s11192-006-0144-7

Groom, C. R., Bruno, I. J., Lightfoot, M. P., & Ward, S. C. (2016). The Cambridge Structural Database. *Acta Crystallographica Section B, 72*(2), 171-179. https://doi.org/10.1107/S2052520616003954

Harzing, A.-W., & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. *Scientometrics, 106*(2), 787-804. https://doi.org/10.1007/s11192-015-1798-9

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output [Article]. *Proceedings of the National Academy of Sciences, 102*(46), 16569-16572. https://doi.org/doi:10.1073/pnas.0507655102

Jung, K. S., Hong, K. W., Jo, H. Y., Choi, J., Ban, H. J., Cho, S. B., & Chung, M. (2020). KRGDB: the large-scale variant database of 1722 Koreans based on whole genome sequencing. *Database: The Journal of Biological Databases and Curation*, 1-7. https://doi.org/10.1093/database/baz146

Kaur, H., Bhalla, S., Kaur, D., & Raghava, G. P. S. (2020). CancerLivER: a database of liver cancer gene expression resources and biomarkers. *Database: The Journal of Biological Databases and Curation*, 1-11. https://doi.org/10.1093/database/baaa012

Kochen, M. (1963). On natural information systems: pragmatic aspects of information retrieval. *Methods of Information in Medicine, 2*(04), 143-147. https://doi.org/10.1055/s-0038-1636217

Krauskopff, M. (1994). Epistemometria, a term contributing to express the meaning and potential methodologies of scientometrics in Spanish speaking countries. *Scientometrics*, *30*(2), 425-428. https://doi.org/10.1007/BF02018117

Luo, J. Y., Wei, C. C., Liu, H. J., Cheng, S. K., Xiao, Y. J., Wang, X. Q., Yan, J. B., & Liu, J. X. (2020, Jun). MaizeCUBIC: a comprehensive variation database for a maize synthetic population. *Database: The Journal of Biological Databases and Curation*, 1-8. https://doi.org/10.1093/database/baaa044

Lyman, P. V., Hal R. (2003). *How much information 2003*. Retrieved Sep from https://groups.ischool.berkeley.edu/archive/how-much-info-2003/

Microsoft Corporation. (2018). *Microsoft Excel*. In (Version 2016, 2019, 365) https://office.microsoft.com/excel

Olle, T. W. (2006). Nineteen Sixties History of Data Base Management. IFIP International Conference on the History of Computing,

Oracle Corporation. (s.f.). *¿Qué es una base de datos?* Retrieved Jul from https://www.oracle.com/cl/database/what-is-database/

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hrobjartsson, A., Lalu, M. M., Li, T. J., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Bmj-British Medical Journal*, *372*, 1-9. https://doi.org/10.1136/bmj.n71

Pritchard, A. (1969). Statistical bibliography or bibliometrics. *Journal of Documentation*, *25*, 348.

RStudio Team. (2020). *RStudio: integrated development for R.* In (Version 4.2.1) RStudio, PBC. http://www.rstudio.com/

Sorokina, M., & Steinbeck, C. (2020). Review on natural products databases: where to find data in 2020. *Journal of Cheminformatics, 12*(1), 1-51. https://doi.org/10.1186/s13321-020-00424-9

Spinak, E. (1996). *Diccionario Enciclopédico de Bibliometría, Cienciometría e Informetría*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000243329

Spoor, S., Wytko, C., Soto, B., Chen, M., Almsaeed, A., Condon, B., Herndon, N., Hough, H., Jung, S., Staton, M., Wegrzyn, J., Main, D., Feltus, F. A., & Ficklin, S. P. (2020). Tripal and Galaxy: supporting reproducible scientific workflows for community biological databases. *Database: The Journal of Biological Databases and Curation*, 1-9. https://doi.org/10.1093/database/baaa032

Suarez Colorado, Y., & Pérez-Anaya, O. (2018). La evaluación de la actividad científica: Indicadores bibliométricos. In J. H. Ávila Toscano (Ed.), *Cienciometría y bibliometría. El estudio de la producción científica: Métodos, enfoques y aplicaciones en el estudio de las Ciencias Sociales* (pp. 96-118). Corporación Universitaria Reformada. https://dialnet.unirioja.es/descarga/articulo/6652726.pdf

The EndNote Team. (2013). *EndNote.* In (Version EndNote X9) [64 bit]. Clarivate Analytics.

Wanyama, S. B., McQuaid, R. W., & Kittler, M. (2022). Where you search determines what you find: the effects of bibliographic databases on systematic reviews. *International Journal of Social Research Methodology, 25*(3), 409-422. https://doi.org/10.1080/13645579.2021.1892378

ANEXOS

- **Tabla 25.** Registros con igual título sin edición. Resalta las filas con problemas por mal conteo de citas de WoS
- **Tabla 26.** Registros duplicados ordenados por grupo. Muestra principales diferencias entre ambos grupos.
- **Tabla 27.** Registros duplicados ordenados por pares. Indica los registros conservados y destaca los cambios por unificación del conteo de citas.

Tabla 25. Registros con igual título sin edición. Resalta las filas con problemas por mal conteo de citas de WoS

Authors*	Article Title	Source Title	TC, WoS CC**	TC, All DB***	Publication Year	Volume	DOI	UT (Unique WOS ID)
Sayers, EW; Barrett, T; Benson, DA; Bolton, E; (…)	**Database resources of the National Center for Biotechnology Information**	NUCLEIC ACIDS RESEARCH	**2530**	**2563**	**2012**	40	10.1093/nar/gkr1184	WOS:000298601300003
Acland, A; Agarwala, R; Barrett, T; Beck, J; Benson, DA; (…)	**Database resources of the National Center for Biotechnology Information**	NUCLEIC ACIDS RESEARCH	**2530**	**2563**	**2013**	41	10.1093/nar/gks1189	WOS:000312893300002
Acland, A; Agarwala, R; Barrett, T; Beck, J; Benson, DA; (…)	Database resources of the National Center for Biotechnology Information	NUCLEIC ACIDS RESEARCH	286	293	2014	42	10.1093/nar/gkt1146	WOS:000331139800002
Agarwala, R; Barrett, T; Beck, J; Benson, DA; (…)	Database resources of the National Center for Biotechnology Information	NUCLEIC ACIDS RESEARCH	172	173	2015	43	10.1093/nar/gku1130	WOS:000350210400002
Agarwala, R; Barrett, T; Beck, J; Benson, DA; (…)	Database resources of the National Center for Biotechnology Information	NUCLEIC ACIDS RESEARCH	228	231	2016	44	10.1093/nar/gkv1290	WOS:000371261700002
Agarwala, R; Barrett, T; Beck, J; Benson, DA; (…)	**Database resources of the National Center for Biotechnology Information**	NUCLEIC ACIDS RESEARCH	**2530**	**2563**	**2018**	46	10.1093/nar/gkx1095	WOS:000419550700002
Sayers, EW; Agarwala, R; Bolton, EE; Brister, (…)	**Database resources of the National Center for Biotechnology Information**	NUCLEIC ACIDS RESEARCH	**2530**	**2563**	**2019**	47	10.1093/nar/gky1069	WOS:000462587400004
Caspi, R; Altman, T; Dreher, K; Fulcher, CA; Subhraveti, (…)	The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases	NUCLEIC ACIDS RESEARCH	431	433	2012	40	10.1093/nar/gkr1014	WOS:000298601300112
Caspi, R; Altman, T; Billington, R; Dreher, K; (…)	The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases	NUCLEIC ACIDS RESEARCH	610	615	2014	42	10.1093/nar/gkt1103	WOS:000331139800069
Caspi, R; Billington, R; Ferrer, L; Foerster, H; (…)	The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases	NUCLEIC ACIDS RESEARCH	574	581	2016	44	10.1093/nar/gkv1164	WOS:000371261700065

* Para ver todos los autores ir a la fuente original; ** Times Cited, WoS Core; *** Times Cited, All Databases

Tabla 26. Registros duplicados ordenados por grupo. Muestra principales diferencias entre ambos grupos.

Authors[*]	*Article Title*	*TC, All DB*[**]	*Publication Date*	*Publication Year*	*Volume*	DOI	*UT (Unique WOS ID)*	TC [***]
Richardson, KE; (…)	RNA CoSSMos 2.0: an improved searchable (…)	1	JAN 16	2020		10.1093/database/baz153	WOS:000511355900001	**28**
Mei, SQ; Huang (…)	GREG-studying transcriptional regulation us(…)	0	FEB 13	2020		10.1093/database/baz162	WOS:000519579700001	
Jung, KS; Hong, (…)	KRGDB: the large-scale variant database of (…)	11	MAR 4	2020		10.1093/database/baz146	WOS:000521177500001	
Kaur, H; Bhalla, (…)	CancerLivER: a database of liver cancer gen(…)	3	MAR 7	2020		10.1093/database/baaa012	WOS:000521180200001	
Czipa, E; Schille (…)	ChIPSummitDB: a ChIP-seq-based database (…)	4	JAN 14	2020		10.1093/database/baz141	WOS:000521180700001	
de Souza, EDF; H(…)	Ewe: a web-based ethnobotanical database(…)	0	FEB 12	2020		10.1093/database/baz144	WOS:000521183100001	
Zhao, L; Wu, XH(…)	ctcRbase: the gene expression database of c(…)	3	APR 15	2020		10.1093/database/baaa020	WOS:000527712000001	
Du, LM; Guo, T; (…)	MACSNVdb: a high-quality SNV database fo(…)	0	MAY 4	2020		10.1093/database/baaa027	WOS:000530704300001	
Luo, JY; Wei, CC (…)	MaizeCUBIC: a comprehensive variation dat(…)	0	JUN 16	2020		10.1093/database/baaa044	WOS:000545997600001	
Spoor, S; Wytko(…)	Tripal and Galaxy: supporting reproducible (…)	0	JUL 4	2020		10.1093/database/baaa032	WOS:000548859200001	
Kim, JH; Park, S (…)	Gliome database: a comprehensive web-ba(…)	0	JUL 31	2020		10.1093/database/baaa057	WOS:000562690000001	
Wang, XR; Liu, Z(…)	SPDB: a specialized database and web-base(…)	0	AUG 6	2020		10.1093/database/baaa063	WOS:000562695400001	
Wanichthanarak(…)	ThRSDB: a database of Thai rice starch com(…)	0	DEC 1	2020		10.1093/database/baaa068	WOS:000597219900001	
Tao, YT; Ding, XB(…)	Predicted rat interactome database and ge(…)	0	NOV 20	2020		10.1093/database/baaa086	WOS:000597229000001	
Wang, JR; Fu, W(…)	WGVD: an integrated web-database for wh(…)	0	NOV 11	2020		10.1093/database/baaa090	WOS:000597230700001	
Tan, C; Chapman(…)	BarleyVarDB: a database of barley genomic (…)	5	NOV 28	2020		10.1093/database/baaa091	WOS:000597231900001	
Mochao, H; Bara(…)	KiMoSys 2.0: an upgraded database for sub(…)	1	NOV 28	2020		10.1093/database/baaa093	WOS:000597232700001	
Zhou, YK; Chen, (…)	mPPI: a database extension to visv-ilize stru(…)	0	JUN 22	2021		10.1093/database/baab036	WOS:000687616900001	
Zhou, YK; Chen, (…)	mPPI: a database extension to visualize stru(…)	0	JUN 22	2021	2021	10.1093/database/baab036	WOS:000766827400004	**5**
Wanichthanarak(…)	ThRSDB: a database of Thai rice starch com(…)	0	DEC 1	2020	2020	10.1093/database/baaa068	WOS:000776694500004	
Tan, C; Chapman(…)	BarleyVarDB: a database of barley genomic (…)	0	NOV 28	2020	2020	10.1093/database/baaa091	WOS:000776694900003	
Mochao, H; Bara(…)	KiMoSys 2.0: an upgraded database for sub(…)	0	NOV 28	2020	2020	10.1093/database/baaa093	WOS:000776694900006	
Tao, YT; Ding, XB(…)	Predicted rat interactome database and ge(…)	0	NOV 20	2020	2020	10.1093/database/baaa086	WOS:000776695100001	
Spoor, S; Wytko(…)	Tripal and Galaxy: supporting reproducible (…)	1	JAN 1	2020	2020	10.1093/database/baaa032	WOS:000776695200016	
Kaur, H; Bhalla, (…)	CancerLivER: a database of liver cancer gen(…)	1	JAN 1	2020	2020	10.1093/database/baaa012	WOS:000776695200020	
de Souza, EDF; H(…)	Ewe: a web-based ethnobotanical database(…)	0	JAN 1	2020	2020	10.1093/database/baz144	WOS:000776695200021	
Zhao, L; Wu, XH(…)	ctcRbase: the gene expression database of c(…)	0	JAN 1	2020	2020	10.1093/database/baaa020	WOS:000776695200025	
Kim, JH; Park, S (…)	Gliome database: a comprehensive web-ba(…)	0	JAN 1	2020	2020	10.1093/database/baaa057	WOS:000776695200037	
Jung, KS; Hong, (…)	KRGDB: the large-scale variant database of (…)	2	JAN 1	2020	2020	10.1093/database/baz146	WOS:000776695200043	
Richardson, KE; (…)	RNA CoSSMos 2.0: an improved searchable (…)	0	JAN 1	2020	2020	10.1093/database/baz153	WOS:000776695200045	
Luo, JY; Wei, CC (…)	MaizeCUBIC: a comprehensive variation dat(…)	1	JAN 1	2020	2020	10.1093/database/baaa044	WOS:000776695200050	
Wang, XR; Liu, Z(…)	SPDB: a specialized database and web-base(…)	0	JAN 1	2020	2020	10.1093/database/baaa063	WOS:000776695200055	
Du, LM; Guo, T; (…)	MACSNVdb: a high-quality SNV database fo(…)	0	JAN 1	2020	2020	10.1093/database/baaa027	WOS:000776695200060	
Mei, SQ; Huang (…)	GREG-studying transcriptional regulation us(…)	0	JAN 1	2020	2020	10.1093/database/baz162	WOS:000776695200073	
Czipa, E; Schille (…)	ChIPSummitDB: a ChIP-seq-based database (…)	0	JAN 1	2020	2020	10.1093/database/baz141	WOS:000776695200084	
Wang, JR; Fu, W(…)	WGVD: an integrated web-database for wh(…)	0	JAN 1	2020	2020	10.1093/database/baaa090	WOS:000776695200097	

[*] Para ver todos los autores ir a la fuente original; [**] *Times Cited, All Databases;* [***] Total de citas para el grupo de registros.

Tabla 27. Registros duplicados ordenados por pares. Indica los registros conservados y destaca los cambios por unificación del conteo de citas.

Doc. Type	*Authors*[*]	*Article Title*	*TC, All DB*[**]	*Publication Date*	*Public. Year*	*Volume*	DOI	*UT (Unique WOS ID)*	Acción	Dupl.[***]	TC (c)[****]
Article	Tan, C; Chapman, B; Wang, PH; Zhang, QS; Zhou, GF; Zhang, (…)	BarleyVarDB: a database of barley genomic variation	5	NOV 28	2020		10.1093/database/baaa091	WOS:000597231900001	Guardado	DOI, TI, AU	**5**
Article	Tan, C; Chapman, B; Wang, PH; Zhang, QS; Zhou, GF; Zhang, (…)	BarleyVarDB: a database of barley genomic variation	0	NOV 28	2020	2020	10.1093/database/baaa091	WOS:000776694900003	Descartado		
Article	Kaur, H; Bhalla, S; Kaur, D; Raghava, GPS	CancerLivER: a database of liver cancer gene expression resources and biomarkers	3	MAR 7	2020		10.1093/database/baaa012	WOS:000521180200001	**Guardado y modificado**	DOI, TI, AU	**4**
Article	Kaur, H; Bhalla, S; Kaur, D; Raghava, GPS	CancerLivER: a database of liver cancer gene expression resources and biomarkers	1	JAN 1	2020	2020	10.1093/database/baaa012	WOS:000776695200020	Descartado		
Article	Czipa, E; Schiller, M; Nagy, T; Kontra, L; Steiner, L; Koller, J; Palne-Szen, O; Barta, E	ChIPSummitDB: a ChIP-seq-based database of human transcription factor binding sites and the topological arra(…)	4	JAN 14	2020		10.1093/database/baz141	WOS:000521180700001	Guardado	DOI, TI, AU	**4**
Article	Czipa, E; Schiller, M; Nagy, T; Kontra, L; Steiner, L; Koller, J; Palne-Szen, O; Barta, E	ChIPSummitDB: a ChIP-seq-based database of human transcription factor binding sites and the topological arra(…)	0	JAN 1	2020	2020	10.1093/database/baz141	WOS:000776695200084	Descartado		
Article	Zhao, L; Wu, XH; Li, T; Luo, J; Dong, D	ctcRbase: the gene expression database of circulating tumor cells and microemboli	3	APR 15	2020		10.1093/database/baaa020	WOS:000527712000001	Guardado	DOI, TI, AU	**3**
Article	Zhao, L; Wu, XH; Li, T; Luo, J; Dong, D	ctcRbase: the gene expression database of circulating tumor cells and microemboli	0	JAN 1	2020	2020	10.1093/database/baaa020	WOS:000776695200025	Descartado		
Article	de Souza, EDF; Hawkins, JA	Ewe: a web-based ethnobotanical database for storing and analysing data	0	FEB 12	2020		10.1093/database/baz144	WOS:000521183100001	Guardado	DOI, TI, AU	**0**
Article	de Souza, EDF; Hawkins, JA	Ewe: a web-based ethnobotanical database for storing and analysing data	0	JAN 1	2020	2020	10.1093/database/baz144	WOS:000776695200021	Descartado		
Article	Kim, JH; Park, SH; Han, J; Ko, PW; Kwon, D; Suk, K	Gliome database: a comprehensive web-based tool to access and analyze glia secretome data	0	JUL 31	2020		10.1093/database/baaa057	WOS:000562690000001	Guardado	DOI, TI, AU	**0**
Article	Kim, JH; Park, SH; Han, J; Ko, PW; Kwon, D; Suk, K	Gliome database: a comprehensive web-based tool to access and analyze glia secretome data	0	JAN 1	2020	2020	10.1093/database/baaa057	WOS:000776695200037	Descartado		
Article	Mei, SQ; Huang, XW; Xie, CS; Mora, A	GREG-studying transcriptional regulation using integrative graph databases	0	FEB 13	2020		10.1093/database/baz162	WOS:000519579700001	Guardado	DOI, TI, AU	**0**
Article	Mei, SQ; Huang, XW; Xie, CS; Mora, A	GREG-studying transcriptional regulation using integrative graph databases	0	JAN 1	2020	2020	10.1093/database/baz162	WOS:000776695200073	Descartado		

Doc. Type	*Authors**	*Article Title*	*TC, All DB***	*Publication Date*	*Public. Year*	*Volume*	DOI	*UT (Unique WOS ID)*	Acción	Dupl.****	TC (c)*****
Article	Mochao, H; Barahona, P; Costa, RS	KiMoSys 2.0: an upgraded database for submitting, storing and accessing experimental data for kinetic modeling	1	NOV 28	2020		10.1093/database/baaa093	WOS:000597232700001	Guardado	DOI, TI, AU	**1**
Article	Mochao, H; Barahona, P; Costa, RS	KiMoSys 2.0: an upgraded database for submitting, storing and accessing experimental data for kinetic modeling	0	NOV 28	2020	2020	10.1093/database/baaa093	WOS:000776694900006	Descartado		
Article	Jung, KS; Hong, KW; Jo, HY; Choi, J; Ban, HJ; Cho, SB; Chung, M	KRGDB: the large-scale variant database of 1722 Koreans based on whole genome sequencing	11	MAR 4	2020		10.1093/database/baz146	WOS:000521177500001	**Guardado y modificado**	DOI, TI, AU	**14***
Article	Jung, KS; Hong, KW; Jo, HY; Choi, J; Ban, HJ; Cho, SB; Chung, M	KRGDB: the large-scale variant database of 1722 Koreans based on whole genome sequencing	**2**	JAN 1	2020	2020	10.1093/database/baz146	WOS:000776695200043	Descartado		
Correction	Jung, KS; Hong, KW; Jo, HY; Choi, J; Ban, HJ; Cho, SB; Chung, M	KRGDB: the large-scale variant database of 1722 Koreans based on whole genome sequencing **(vol 2020, baz146, 2020)**	**3**	JAN 1	2020	2020	10.1093/database/baaa030	WOS:000776695200093	No aplica		
Correction	Jung, KS; Hong, KW; Jo, HY; Choi, J; Ban, HJ; Cho, SB; Chung, M	KRGDB: the large-scale variant database of 1722 Koreans based on whole genome sequencing **(vol 2020, baz146, 2020)**	0	APR 29	2020		10.1093/database/baaa030	WOS:000530366200001	No aplica		
Article	Du, LM; Guo, T; Liu, Q; Li, J; Zhang, XY; Xing, JC; Yue, BS; Fan, ZX	MACSNVdb: a high-quality SNV database for interspecies genetic divergence investigation among macaques	0	MAY 4	2020		10.1093/database/baaa027	WOS:000530704300001	Guardado	DOI, TI	**0**
Article	Du, LM; Guo, T; Liu, Q; Li, J; Zhang, XY; Xing, JC; Yue, BS; **Li, J**; Fan, ZX	MACSNVdb: a high-quality SNV database for interspecies genetic divergence investigation among macaques	0	JAN 1	2020	2020	10.1093/database/baaa027	WOS:000776695200060	Descartado		
Article	Luo, JY; Wei, CC; Liu, HJ; Cheng, SK; Xiao, YJ; Wang, XQ; Yan, JB; Liu, JX	MaizeCUBIC: a comprehensive variation database for a maize synthetic population	0	JUN 16	2020		10.1093/database/baaa044	WOS:000545997600001	**Guardado y modificado**	DOI, TI, AU	**1**
Article	Luo, JY; Wei, CC; Liu, HJ; Cheng, SK; Xiao, YJ; Wang, XQ; Yan, JB; Liu, JX	MaizeCUBIC: a comprehensive variation database for a maize synthetic population	1	JAN 1	2020	2020	10.1093/database/baaa044	WOS:000776695200050	Descartado		
Article	Zhou, YK; Chen, HJ; Li, SD; Chen, M	mPPI: a database extension to **visualize** structural interactome in a one-to-many manner	0	JUN 22	2021	2021	10.1093/database/baab036	WOS:000766827400004	Descartadp	DOI, AU	**0**
Article	Zhou, YK; Chen, HJ; Li, SD; Chen, M	mPPI: a database extension to **visv-ilize** structural interactome in a one-to-many manner	0	JUN 22	2021		10.1093/database/baab036	WOS:000687616900001	Guardado		

Doc. Type	*Authors*[*]	*Article Title*	*TC, All DB*[**]	*Publication Date*	*Public. Year*	*Volume*	DOI	*UT (Unique WOS ID)*	Acción	Dupl.[***]	TC (c)[****]
Article	Tao, YT; Ding, XB; Jin, J; Zhang, HB; Guo, WP; Ruan, L; Yang, Q(…)	Predicted rat interactome database and gene set linkage analysis	0	NOV 20	2020		10.1093/database/baaa086	WOS:000597229000001	Guardado	DOI, TI, AU	**0**
Article	Tao, YT; Ding, XB; Jin, J; Zhang, HB; Guo, WP; Ruan, L; Yang, Q(…)	Predicted rat interactome database and gene set linkage analysis	0	NOV 20	2020	2020	10.1093/database/baaa086	WOS:000776695100001	Descartado		
Article	Richardson, KE; Kirkpatrick, CC; Znosko, BM	RNA CoSSMos 2.0: an improved searchable database of secondary structure motifs in RNA three-dimens(…)	1	JAN 16	2020		10.1093/database/baz153	WOS:000511355900001	Guardado	DOI, TI, AU	**1**
Article	Richardson, KE; Kirkpatrick, CC; Znosko, BM	RNA CoSSMos 2.0: an improved searchable database of secondary structure motifs in RNA three-dimens(…)	0	JAN 1	2020	2020	10.1093/database/baz153	WOS:000776695200045	Descartado		
Article	Wang, XR; Liu, ZB; Li, XY; Li, DW; Cai, JY; Yan, H	SPDB: a specialized database and web-based analysis platform for swine pathogens	0	AUG 6	2020		10.1093/database/baaa063	WOS:000562695400001	Guardado	DOI, TI, AU	**0**
Article	Wang, XR; Liu, ZB; Li, XY; Li, DW; Cai, JY; Yan, H	SPDB: a specialized database and web-based analysis platform for swine pathogens	0	JAN 1	2020	2020	10.1093/database/baaa063	WOS:000776695200055	Descartado		
Article	Wanichthanarak, K; Thitisaksakul, M	ThRSDB: a database of Thai rice starch composition, molecular structure and functionality	0	DEC 1	2020		10.1093/database/baaa068	WOS:000597219900001	Guardado	DOI, TI, AU	**0**
Article	Wanichthanarak, K; Thitisaksakul, M	ThRSDB: a database of Thai rice starch composition, molecular structure and functionality	0	DEC 1	2020	2020	10.1093/database/baaa068	WOS:000776694500004	Descartado		
Article	Spoor, S; Wytko, C; Soto, B; Chen, M; Almsaeed, A; Condon, B; (…)	Tripal and Galaxy: supporting reproducible scientific workflows for community biological databases	0	JUL 4	2020		10.1093/database/baaa032	WOS:000548859200001	**Guardado y modificado**	DOI, TI, AU	**1**
Article	Spoor, S; Wytko, C; Soto, B; Chen, M; Almsaeed, A; Condon, B; (…)	Tripal and Galaxy: supporting reproducible scientific workflows for community biological databases	1	JAN 1	2020	2020	10.1093/database/baaa032	WOS:000776695200016	Descartado		
Article	Wang, JR; Fu, WW; Wang, R; Hu, DX; Cheng, H; Zhao, J; Jiang, Y; Kang, ZS	WGVD: an integrated web-database for wheat genome variation and selective signatures	0	NOV 11	2020		10.1093/database/baaa090	WOS:000597230700001	Guardado	DOI, TI, AU	**0**
Article	Wang, JR; Fu, WW; Wang, R; Hu, DX; Cheng, H; Zhao, J; Jiang, Y; Kang, ZS	WGVD: an integrated web-database for wheat genome variation and selective signatures	0	JAN 1	2020	2020	10.1093/database/baaa090	WOS:000776695200097	Descartado		

[*] Para ver todos los autores ir a la fuente original; [**] *Times Cited, All Databases;* [***] Principales campos duplicados; [****] Total de citas conservados tras deduplicación