

Proyecto de Bases de datos para un análisis bibliométrico

Nubia Fernanda Sánchez Bello¹

¹Facultad de Ingeniería y Ciencias Básicas

Universidad Central

Maestría en Analítica de Datos

Curso de Bases de Datos

Bogotá, Colombia

¹nsanchezb1@ucentral.edu.co

18 de noviembre de 2022

Índice

1. Introducción (Max 250 Palabras) - (<i>Primera entrega</i>)	3
2. Características del proyecto de investigación (<i>Primera entrega</i>)	4
2.1. Título del proyecto de investigación (<i>Primera entrega</i>)	4
2.2. Objetivo general (<i>Primera entrega</i>)	4
2.2.1. Objetivos específicos (<i>Primera entrega</i>)	4
2.3. Alcance (<i>Primera entrega</i>)	4
2.4. Pregunta de investigación (<i>Primera entrega</i>)	5
2.5. Hipótesis (<i>Primera entrega</i>)	5
3. Reflexiones sobre el origen de datos e información (<i>Primera entrega</i>)	6
3.1. ¿Cual es el origen de los datos e información ? (<i>Primera entrega</i>) .	6
3.2. ¿Cuales son las consideraciones legales o eticas del uso de la información? (<i>Primera entrega</i>)	6
3.3. ¿Cuales son los retos de la información y los datos que utilizara en la base de datos en terminos de la calidad y la consolidación? (<i>Primera entrega</i>)	7
3.4. ¿Que espera de la utilización de un sistema de Bases de Datos para su proyecto? (<i>Primera entrega</i>)	7

4. Diseño del Modelo de Datos del SMBD (Sistema Manejador de Bases de Datos)(Primera entrega)	8
4.1. Características del SMBD (Sistema Manejador de Bases de Datos) para el proyecto (<i>Primera entrega</i>)	8
4.2. Diagrama modelo de datos (<i>Primera entrega</i>)	8
4.3. Imágenes de la Base de Datos (<i>Primera entrega</i>)	9
4.4. Código SQL - lenguaje de definición de datos (DDL) (<i>Primera entrega</i>)	10
4.5. Código SQL - Manipulación de datos (DML) (<i>Primera entrega</i>) . .	11
4.6. Código SQL + Resultados: Vistas (<i>Primera entrega</i>)	14
4.7. Código SQL + Resultados: Triggers (<i>Primera entrega</i>)	16
4.8. Código SQL + Resultados: Funciones (<i>Primera entrega</i>)	17
4.9. Código SQL + Resultados: procedimientos almacenados (<i>Primera entrega</i>)	18
5. Bases de Datos No-SQL (Segunda entrega)	19
5.1. Diagrama Bases de Datos No-SQL (<i>Segunda entrega</i>)	19
5.2. SMBD utilizado para la Base de Datos No-SQL (<i>Segunda entrega</i>)	19
6. Bibliografía	20

1. Introducción (Max 250 Palabras) - (*Primera entrega*)

Si consideramos que la utilidad de la ciencia y las investigaciones dependen de si los resultados son divulgados de una manera eficaz, esta utilidad involucra también a los artículos científicos; al ser productos de divulgación, un artículo es de utilidad si transmite su conocimiento. Este último aspecto suele intentarse medir de forma objetiva a través de los indicadores de citación, partiendo del supuesto de que el dato de qué tan citado es un artículo, refleja de forma transparente y directa su impacto sobre el conocimiento científico, llegando incluso a considerar que, mientras más alto sea el valor del indicador, mayor relevancia tendrá un artículo y la revista en la que este ha sido publicado (Martinovich, 2020).

Los indicadores de citación, como su nombre indica, se basan en el conteo y análisis de las citaciones que reciben los artículos de revistas científicas, estos análisis se han convertido en una estrategia de las universidades y otras instituciones interesadas en investigación para medir producción y relevancia científica de la misma (Molina-Molina et al., 2020), estos indicadores y otros más se emplean para realizar análisis bibliométricos. Los análisis bibliométricos, principalmente aquellos relacionados con revistas científicas indexadas, han tenido un auge reciente, debido en gran medida al avance y disponibilidad del software capaz de realizar estos análisis, y al acceso a bases de datos de las cuales se puede obtener una gran cantidad de información de forma sencilla (Donthu et al., 2021); adicionalmente, se ha encontrado utilidad en los análisis bibliométricos como fuente de información científica y estrategia para producir investigación de alto impacto.

Los indicadores de citación buscan en sí medir un solo aspecto de la investigación, que es su impacto o utilidad, pero no toda citación es siempre positiva, y desde luego, existen varios factores extrínsecos, no relacionados con calidad o contenido, que pueden afectar la citación de un artículo (Onodera y Yoshikane, 2015). Dentro de los principales factores que afectan la citación se encuentran principalmente su accesibilidad, su diseminación y la autoridad científica de los autores, sin embargo, existen muchos otros que pueden afectar la citación como publicaciones previas de los autores, relación del artículo con otros trabajos relevantes, tendencias científicas, obsolescencia de los resultados, calidad de los aspectos formales, el contexto teórico del artículo, y el tipo de trabajo publicado (Repiso et al., 2021). Con estos cuestionamientos resulta dudoso considerar los indicadores de citación como referentes de la calidad o de la producción científica.

Dentro de las áreas temáticas de las revistas científicas, un área que tiene un interés particular es el área de las ciencias biomédicas, área encargada de los temas relacionados con ciencias de la salud (Navarrete y Pérez, 2019). Este tipo de publicaciones se han convertido en el principal canal de comunicación para la comunidad de esta área (Navarrete y Pérez, 2019), con varias de ellas llegando a convertirse en referentes para la atención de pacientes en tan solo cuestión de días, para bien o para mal; situación claramente reflejada durante la atención de la pandemia por COVID-19, cuando algunos artículos publicados

en revistas consideradas de alta calidad eran tenidos en cuenta rápidamente como referentes, asumidos por la comunidad como una pauta de atención, y posteriormente desestimados por fallas en su elaboración (Anderson et al., 2021). Este es un caso en el cual la alta citación pudo no reflejar la calidad, y ocasionar además dificultades al haber diseminado un conocimiento inadecuado.

2. Características del proyecto de investigación (*Primera entrega*)

2.1. Título del proyecto de investigación (*Primera entrega*)

Identificación de factores que influyen sobre la citación de los artículos de las revistas biomédicas colombianas utilizando algoritmos de Machine learning.

2.2. Objetivo general (*Primera entrega*)

Analizar los factores que influyen sobre la citación de los artículos de las revistas biomédicas colombianas indexadas en Scopus, por medio de algoritmos de Machine learning.

2.2.1. Objetivos específicos (*Primera entrega*)

- Extraer datos correspondientes a los artículos de las revistas biomédicas colombianas reconocidas por Pubindex e indexadas en Scopus.
- Examinar el comportamiento bibliométrico de cada revista científica biomédica y compararlo.
- Identificar los principales factores y tendencias que pueden tener influencia sobre la citación de un artículo científico.
- Categorizar y agrupar las principales temáticas y factores presentados por las revistas científicas biomédicas.
- Describir las posibles correlaciones entre las características de los artículos de las revistas biomédicas.
- Analizar la calidad de revisiones sistemáticas presentes en cada revista biomédica a través del riesgo de sesgo identificado por RobotReviewer

2.3. Alcance (*Primera entrega*)

El alcance de este proyecto es correlacional. La información que será obtenida debe ser descrita para luego ser categorizada y evaluada; se espera entonces obtener una serie de comparaciones que den cuenta del desarrollo que ha tenido la producción científica de las revistas científicas biomédicas colombianas indexadas en Scopus. Una vez sea analizada, la información obtenida permitirá plantear algunas hipótesis y definir algunos vacíos que requerirán mayor investigación en un futuro, sin embargo, empleando los ensayos clínicos, se espera establecer una

relación entre las características de algunos artículos, su calidad y el número de citaciones obtenidas.

2.4. Pregunta de investigación (*Primera entrega*)

¿Cuáles son los factores que afectan la citación de los artículos de las revistas biomédicas en Colombia?

2.5. Hipótesis (*Primera entrega*)

Los factores o características de los artículos que definen una citación, no se relacionan con su calidad.

3. Reflexiones sobre el origen de datos e información (*Primera entrega*)

Los datos provienen de una base de datos cuya principal función es almacenar abstracts y citaciones, y que además tiene herramientas de visualización y análisis (University of Michigan Library, 2022), lo que la convierte en una fuente ideal para obtener información que permita hacer un análisis bibliométrico, y puede llegar a ser una fuente confiable, sin embargo, se debe tener en cuenta que la información capturada por Scopus proviene de los metadatos producidos por cada revista para cada artículo, por lo tanto, no sería extraño que ocasionalmente se encontrarán algunos errores, o incluso que algunos artículos o revistas no estuvieran completamente disponibles para consulta.

La verificación de la calidad de los datos deberá considerarse como un paso intermedio entre su obtención y su consolidación en una base de datos; esto requerirá además cierto grado de automatización por el volumen de información que será manejado.

Existe una limitación importante en este proyecto y es que sólo se están teniendo en cuenta aquellas revistas que fueron indexadas por Pubindex y por Scopus. Este criterio de inclusión se realiza a conveniencia pues Scopus permite un acceso sencillo a una gran cantidad de metadatos de revistas científicas, y Pubindex indexa revistas colombianas con un mínimo de criterios de calidad, lo que garantiza que al menos, los datos que se obtendrán, podrán ser comparables en su gran mayoría, y aunque no representen a la totalidad de las revistas, si representarán a las revistas de mayor relevancia.

3.1. ¿Cual es el origen de los datos e información ? (*Primera entrega*)

Los datos provienen principalmente de Scopus, base de datos de Elsevier, que a través de su proceso de indexación ha catalogado información de 81 millones de documentos (Elsevier, 2022). El buscador de Scopus permite consultar información acerca de artículos y revistas empleando sus metadatos para realizar consultas específicas, y además genera indicadores de citación de manera periódica a nivel de artículo, revista, autor e institución. Adicionalmente se ha consultado el Índice Bibliográfico de Pubindex para establecer cuáles son las revistas biomédicas indexadas por Minciencias en Colombia.

3.2. ¿Cuales son las consideraciones legales o eticas del uso de la información? (*Primera entrega*)

La información de Pubindex se encuentra disponible para consulta pública, y la información de Scopus puede consultarse a través de una cuenta institucional o de Elsevier. La información bibliométrica proviene de artículos que los autores autorizaron fueran publicados, por lo tanto no revelan información privada o sensible.

3.3. ¿Cuales son los retos de la información y los datos que utilizara en la base de datos en terminos de la calidad y la consolidación? (*Primera entrega*)

El principal reto es el volumen de la información a ser consolidada, su almacenamiento y consulta deben ser óptimos para permitir su análisis. El segundo reto es verificar que los metadatos descargados de Scopus sean comparables y tengan una calidad adecuada para realizar los análisis correspondientes.

3.4. ¿Que espera de la utilización de un sistema de Bases de Datos para su proyecto? (*Primera entrega*)

La gestión efectiva de la información obtenida facilitará el proceso de análisis y reducirá la probabilidad de cometer errores u omisiones; además, almacenar un volumen tan grande de información es más eficaz si se realiza a través de una base de datos. Finalmente, el almacenamiento en la base de datos resguardará la información previniendo que la misma sea borrada o alterada por error.

4. Diseño del Modelo de Datos del SMBD (Sistema Manejador de Bases de Datos) (Primera entrega)

4.1. Características del SMBD (Sistema Manejador de Bases de Datos) para el proyecto (Primera entrega)

El SMBD que se va a emplear es MySQL, es una base de datos relacional ampliamente utilizada y que entre sus ventajas cuenta su alta estabilidad, seguridad y disponibilidad de soporte y tutoriales (Suehring, 2002); adicionalmente es posible su integración con Python, lo que facilita el procesamiento de información para los análisis a realizar. Un beneficio adicional que facilita su usabilidad es que es Open source, por lo que fácilmente se obtiene información y software de apoyo en distintas comunidades (Oracle, 2022).

4.2. Diagrama modelo de datos (Primera entrega)

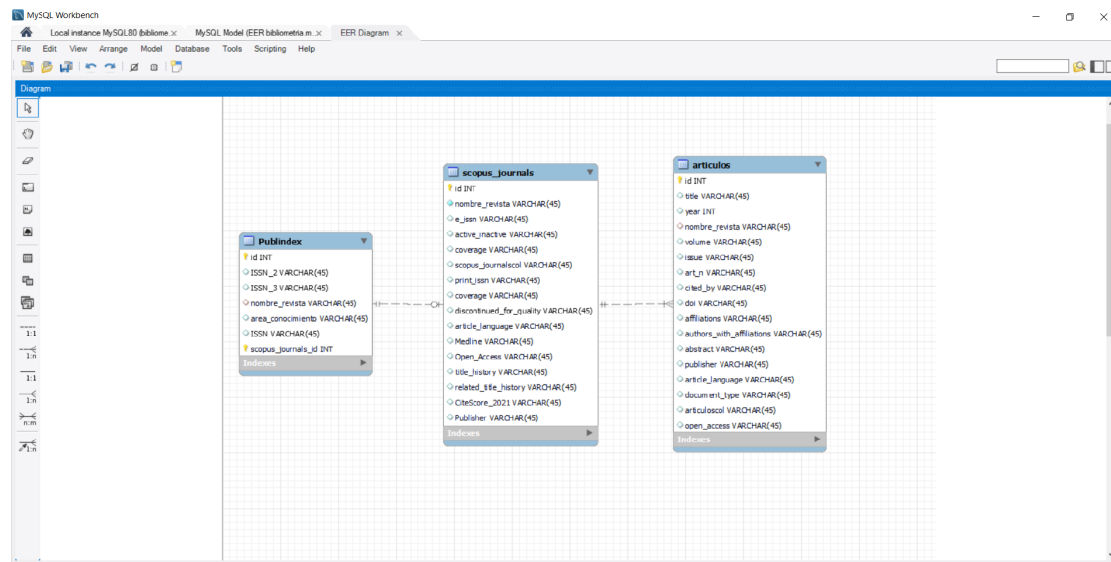


Figura 1: Primera versión del diagrama modelo de datos.

4.3. Imágenes de la Base de Datos (*Primera entrega*)

A continuación se presentan imágenes de las tablas de la base de datos.

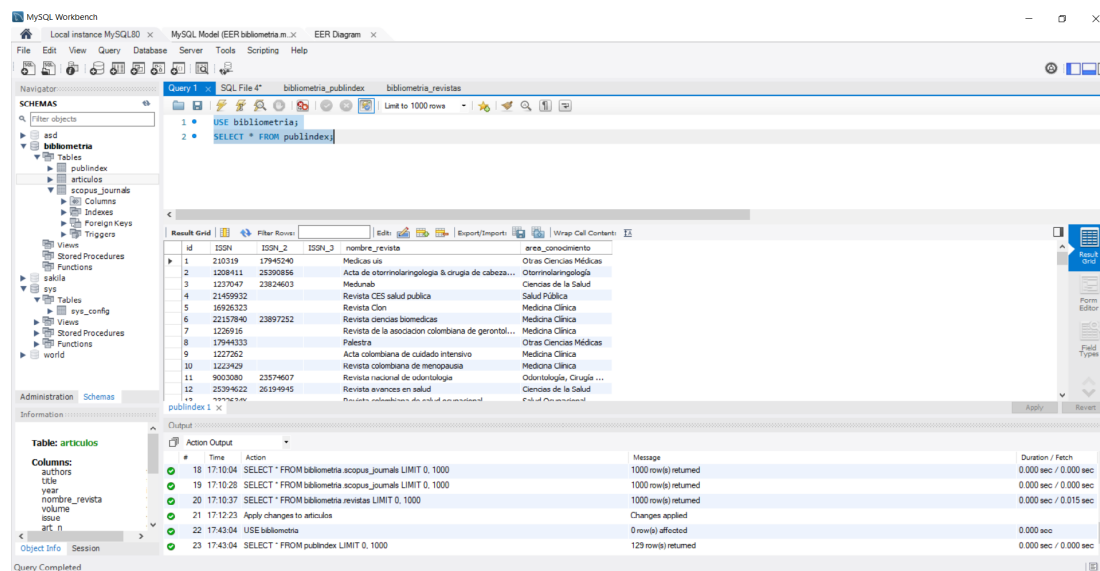


Figura 2: Imagen tabla Publindex.

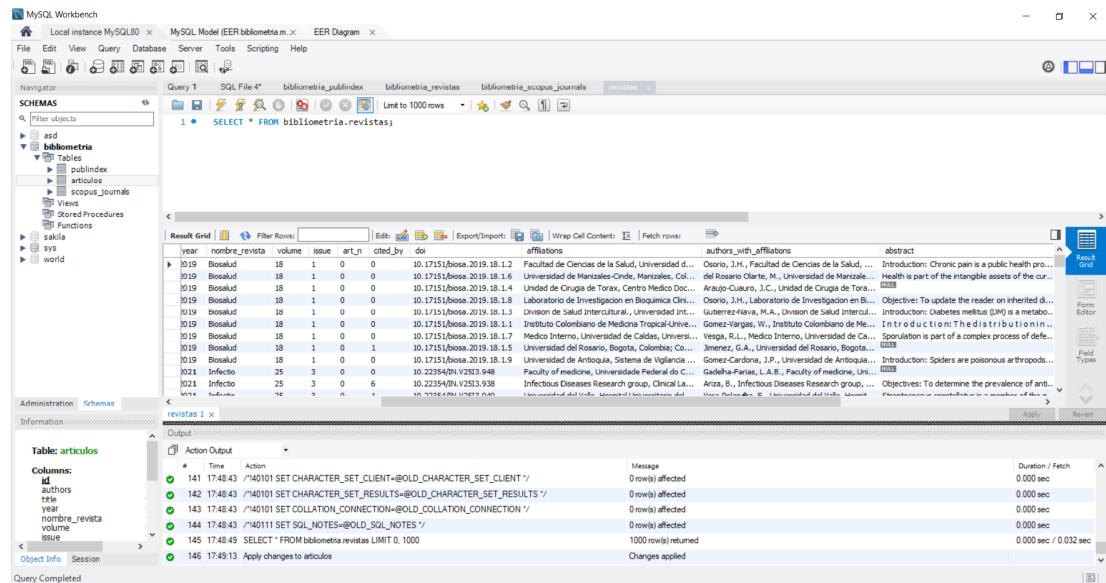


Figura 3: Imagen tabla Artículos.

4.4. Código SQL - lenguaje de definición de datos (DDL) *(Primera entrega)*

En esta sección se encuentra el código SQL empleado para construir las tablas que componen la base. Los datos que se encuentran en cada tabla fueron cargados importando archivos CSV, que es la forma en la que se exporta la consulta de esta información.

```

1
2 create database bibliometria;
3 use bibliometria;
4 create table publindex(
5     id int auto_increment,
6     ISSN int,
7     ISSN_2 varchar(100),
8     ISSN_3 varchar(100),
9     nombre_revista text,
10    area_conocimiento text,
11    primary key(id)
12 );
13
14 create table scopus_journals(
15     id int auto_increment,
16     nombre_revista text,
17     print.ISSN text,
18     e-ISSN text,
19     active_inactive text,
20     coverage text,
21     discontinued_for_quality text,

```

```

22     article_language text,
23     Medline text,
24     Open_Access text,
25     title_history text,
26     related_title_history text,
27     CiteScore_2021 text,
28     Publisher text,
29     primary key(id)
30 );
31
32 create table articulos(
33     id int auto_increment,
34     authors text,
35     title text,
36     year text,
37     nombre_revista text,
38     volume text,
39     issue text,
40     art_n text,
41     cited_by int,
42     doi text,
43     affiliations text,
44     authors_with_affiliations text,
45     abstract text,
46     Publisher text,
47     article_language text,
48     document_type text,
49     Open_Access text,
50     primary key(id)
51 );

```

4.5. Código SQL - Manipulación de datos (DML) (*Primera entrega*)

Se presentan algunos ejemplos de manipulación de datos dentro de las tablas.

En el primer ejemplo se agrega un registro a la tabla Publinindex usando insert:

```

1 insert into publinindex values (130, 99999999, 11111111, 55555555,
2 'Revista Latinoamericana de Anatomia', 'Otras Ciencias Medicas');

```

El último registro de la tabla original tenía el id 129, por tal razón el registro siguiente tendrá id 130, y se colocan valores adecuados para cada columna de la tabla, generando así un nuevo registro.

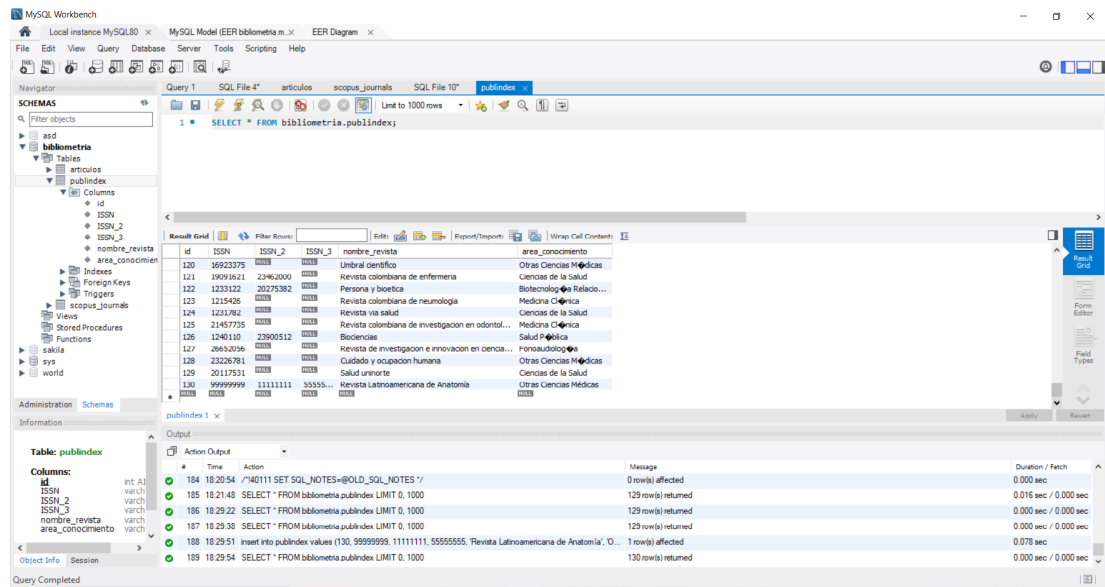


Figura 4: Adición de nuevo registro a tabla Publindex.

Ahora con update se actualiza el ISSN de este registro:

```
1 update publindex set ISSN = 10101010 where id = 130;
```

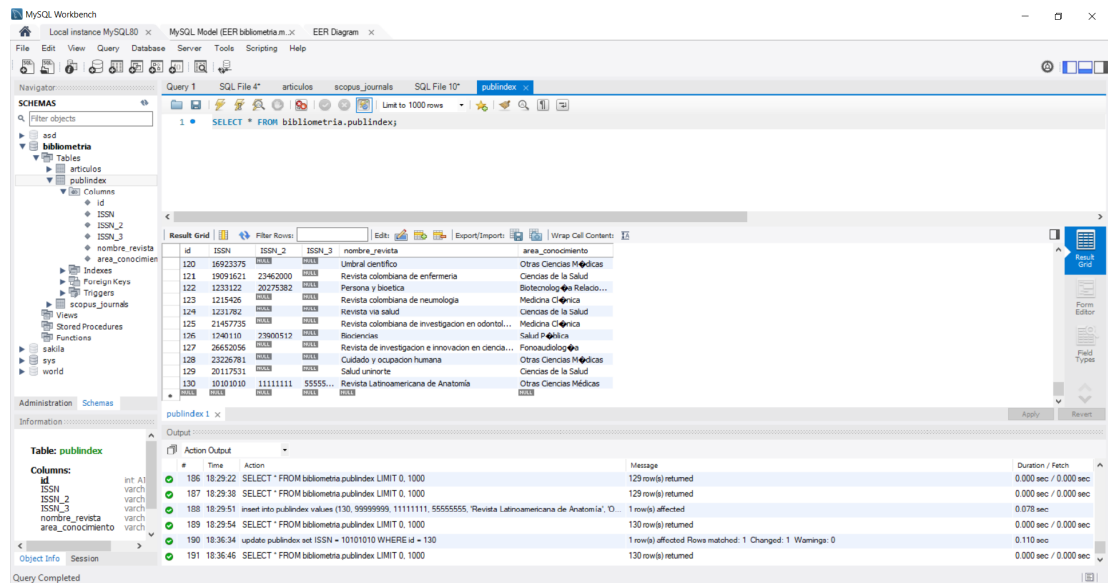


Figura 5: Actualización de registro en tabla Publindex.

Finalmente, con delete se borra ese registro. En mi caso estoy usando “safe update mode”, por lo tanto, MySQL sólo me permite borrar un registro si utilizo la columna que nombré como llave, es decir la columna id:

```
1 delete from publindex where id = 130;
```

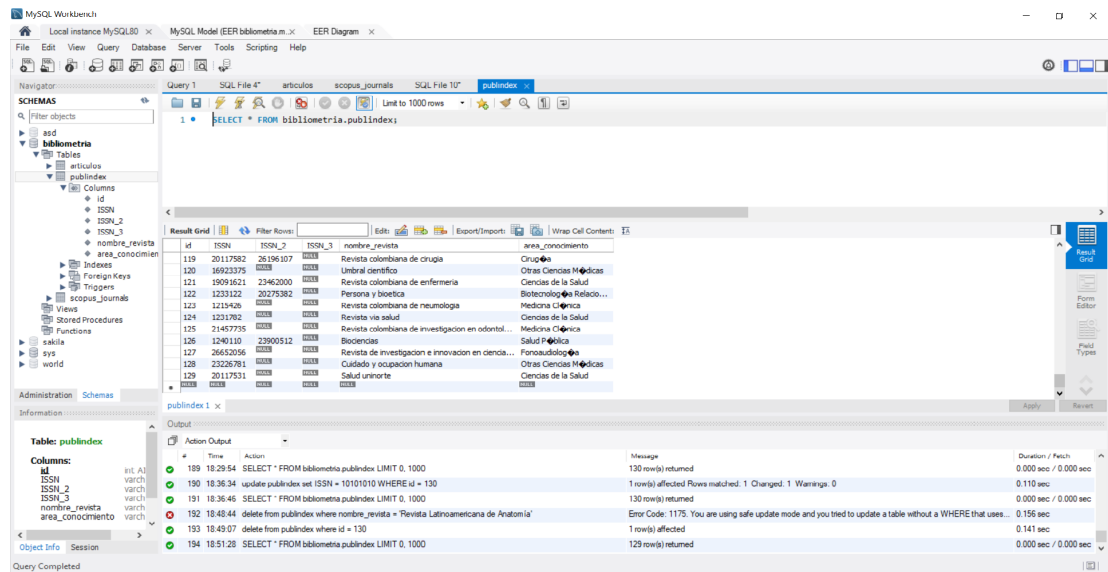


Figura 6: Registro eliminado de la tabla Publindex.

4.6. Código SQL + Resultados: Vistas (*Primera entrega*)

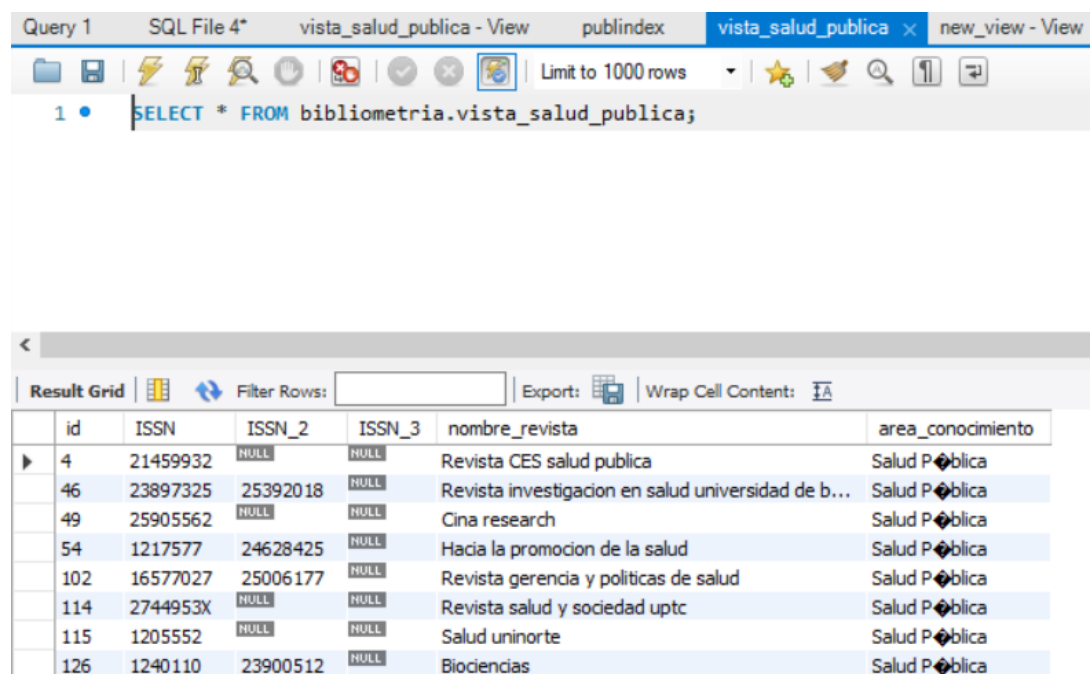
Se presentan algunas vistas de relevancia para análisis de información. En primer lugar una vista que me permita ver revistas de un área en específico, por ejemplo, las revistas colombianas de Salud Pública:

```

1 create view 'vista_salud_publica' as
2 select *
3 from publindex
4 where area_conocimiento = 'Salud Publica';

```

Así puedo ver aquellas revistas indexadas en Publindex cuya área de conocimiento es la Salud Pública:



The screenshot shows a SQL query execution window with the following tabs: Query 1, SQL File 4*, vista_salud_publica - View, publindex, vista_salud_publica, and new_view - View. The query editor contains the following SQL statement:

```
1 • SELECT * FROM bibliometria.vista_salud_publica;
```

The results are displayed in a table with the following columns: id, ISSN, ISSN_2, ISSN_3, nombre_revista, and area_conocimiento. The table contains 8 rows of data, all of which are indexed in Publindex and have 'Salud Pública' as the area of knowledge.

id	ISSN	ISSN_2	ISSN_3	nombre_revista	area_conocimiento
4	21459932	NULL	NULL	Revista CES salud publica	Salud Pública
46	23897325	25392018	NULL	Revista investigacion en salud universidad de b...	Salud Pública
49	25905562	NULL	NULL	Cina research	Salud Pública
54	1217577	24628425	NULL	Hacia la promocion de la salud	Salud Pública
102	16577027	25006177	NULL	Revista gerencia y politicas de salud	Salud Pública
114	2744953X	NULL	NULL	Revista salud y sociedad uptc	Salud Pública
115	1205552	NULL	NULL	Salud uninorte	Salud Pública
126	1240110	23900512	NULL	Biociencias	Salud Pública

Figura 7: Revistas de Salud Pública indexadas en Publindex.

Otra vista de interés es la consulta de cuáles revistas indexadas en Scopus, publican artículos en español:

```
1 create view 'vista_scopus_spa' as
2 select nombre_revista, print.ISSN
3 from scopus_journals
4 where article.language = 'SPA';
```

Query 1 SQL File 4* vista_salud_publica - View pubindex vista_salud_publica vista_scopus_spa - View

Limit to 1000 rows

1 • **SELECT * FROM bibliometria.vista_scopus_spa;**

Result Grid Filter Rows: Export: Wrap Cell Content: [FA](#)

	nombre_revista	print_ISSN
►	Abriu	20148526
	Academia Revista Latinoamericana de Administr...	10128255
	Acotaciones	11307269
	Acta Bioethica	7175906
	Acta Biologica Colombiana	0120548X
	Acta Bioquímica Clínica Latinoamericana	3252957
	Acta Botanica Venezuelica	845906
	Acta Colombiana de Psicología	1239155
	Acta Gastroenterológica Latinoamericana	3009033
	Acta Ginecológica	15776
	Acta Literaria	7160909
	Acta Otorrinolaringológica Española	16519
	Acta Odontológica Española	16540

Figura 8: Revistas de Scopus que publican artículos en español.

4.7. Código SQL + Resultados: Triggers (*Primera entrega*)

Como un trigger de utilidad para esta base, se presenta la opción de guardar un registro de aquellos artículos que sean borrados. Para esto se crea una nueva tabla para almacenar el respaldo:

```

1 use bibliometria;
2 create table articulos_respaldo(
3     id int auto_increment,
4     authors text,
5     title text,
6     year text,
7     nombre_revista text,
8     volume text,
9     issue text,
10    art_n text,
11    cited_by int,
12    doi text,
13    affiliations text,
14    authors_with_affiliations text,
15    abstract text,

```



```

16 Publisher text,
17 article_language text,
18 document_type text,
19 Open_Access text,
20 primary key(id)
21 );

```

y el trigger correspondiente:

```

1 delimiter //
2 create trigger articulo_borrado before delete on articulos
3 for each row begin
4   insert into articulos_respaldo
5   select * from articulos where id=old.id;
6 end //
7 delimiter ;

```

Con el trigger elaborado, si borro un registro, por ejemplo el tercer registro de mi tabla de artículos, este quedará almacenado en la tabla de respaldo:

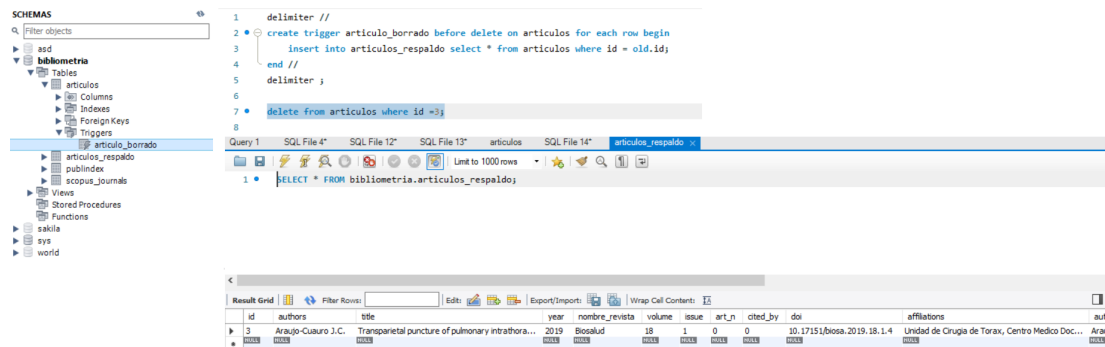


Figura 9: Resultados del trigger para almacenar artículos borrados.

4.8. Código SQL + Resultados: Funciones (*Primera entrega*)

Es de interés revisar cuántos artículos se producen por año, para esto se elabora la siguiente función:

```

1 create function 'publication_year' (numero int)
2 returns integer
3 begin
4   declare numerog int;
5   select count(*) into numerog
6   from articulos where year like 'numero%';
7   return numerog;
8 end

```

La función resultante retorna el número de artículos que fueron publicados en el año empleado con la función:

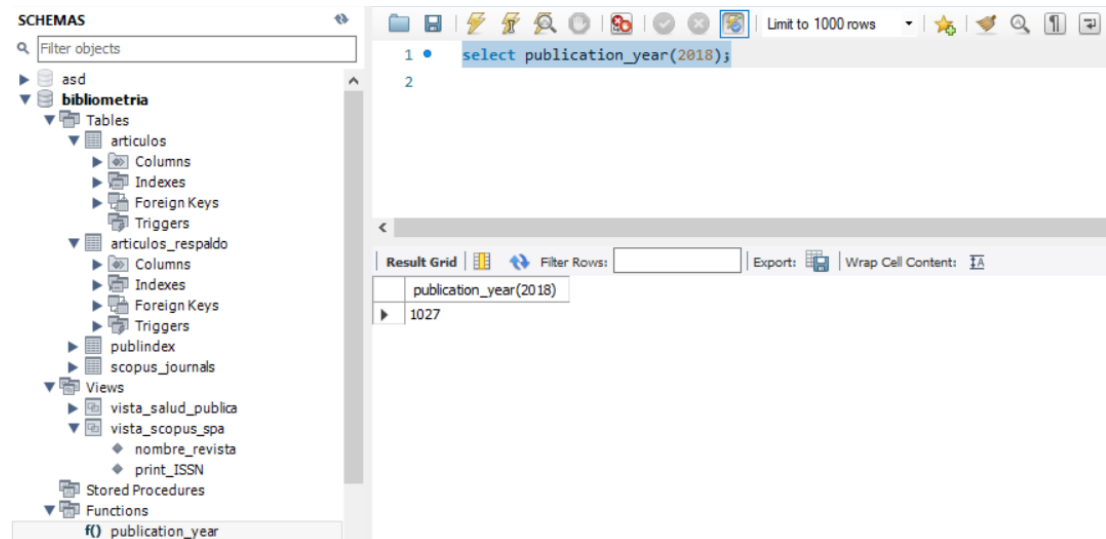


Figura 10: Número de artículos publicados en el año 2018.

4.9. Código SQL + Resultados: procedimientos almacenados (*Primera entrega*)

Una de las consultas principales que deben hacerse en la base de datos es la de revisar cuáles revistas están indexadas tanto en Pubindex como en Scopus. La información de indexación de Pubindex se actualiza anualmente, sin embargo la de Scopus se actualiza con mayor frecuencia, y por eso esta consulta es esencial para definir cuáles revistas se deben incluir dentro del análisis. Para facilitar dicha consulta, se elabora un procedimiento:

```
1 create procedure 'indexados' ()
2 begin
3     select pubindex.nombre_revista , pubindex.issn
4     from pubindex
5     inner join scopus_journals
6     on pubindex.nombre_revista=scopus_journals.nombre_revista;
7 end
```

The screenshot shows a database management interface. On the left, a 'SCHEMAS' pane lists various databases including 'bibliometria' and 'indexados'. The 'indexados' database is selected. The main window displays a query result for 'CALL indexados;'. The result is shown in a 'Result Grid' with two columns: 'nombre_revista' and 'issn'. The data includes various medical journals and their corresponding ISSN numbers.

nombre_revista	issn
Aquichan	16575997
Archivos de medicina	1657320X
Biomedica	1204157
Biosalud	16579550
Colombia medica	16579534
Hacia la promocion de la salud	1217577
Iatreia	1210793
Infectio	1239392
Investigacion y educacion en enfermeria	1205307
Medicina	1205498
Medicina	16920880
Medicina	1205498
Medicina	1205498

Figura 11: Revistas indexadas en Publiindex y en Scopus.

5. Bases de Datos No-SQL (*Segunda entrega*)

5.1. Diagrama Bases de Datos No-SQL (*Segunda entrega*)

Los sistemas No-SQL no son relacionales, por lo tanto, no necesitan un esquema para describir la estructura de los datos almacenados ;sin embargo, aunque no se requiera un esquema, los datos tienden a tener un esquema subyacente, implícito dentro de sus características (Hernández Chillón et al., 2017). Algunos autores han hecho aproximaciones teóricas al respecto, se presenta una estructura de árbol, una de las estrategias de visualización sugerida por un grupo de la Universidad de Murcia (Hernández Chillón et al., 2017):

5.2. SMBD utilizado para la Base de Datos No-SQL (*Segunda entrega*)

Como SMBD para la Base de Datos No-SQL, se propone utilizar Mongo DB, pues su uso es gratuito, ampliamente difundido y con capacidad suficiente para el volumen de información que será manejado. A continuación un ejemplo de una parte de la información cargada en Mongo DB Atlas:

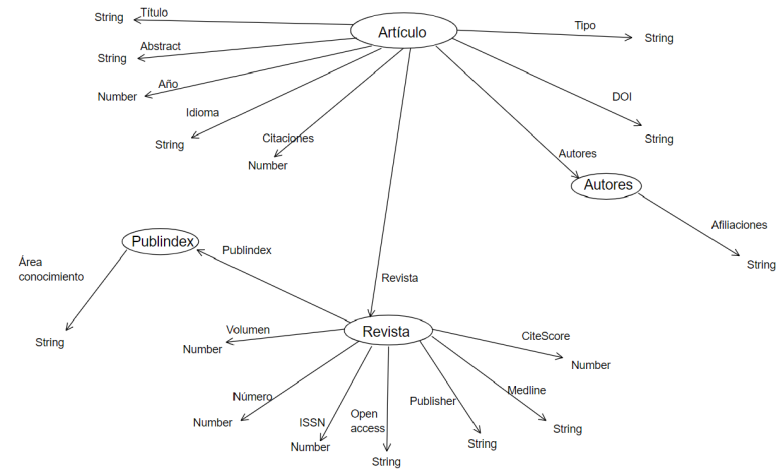


Figura 12: Estructura de árbol que indica las entidades.

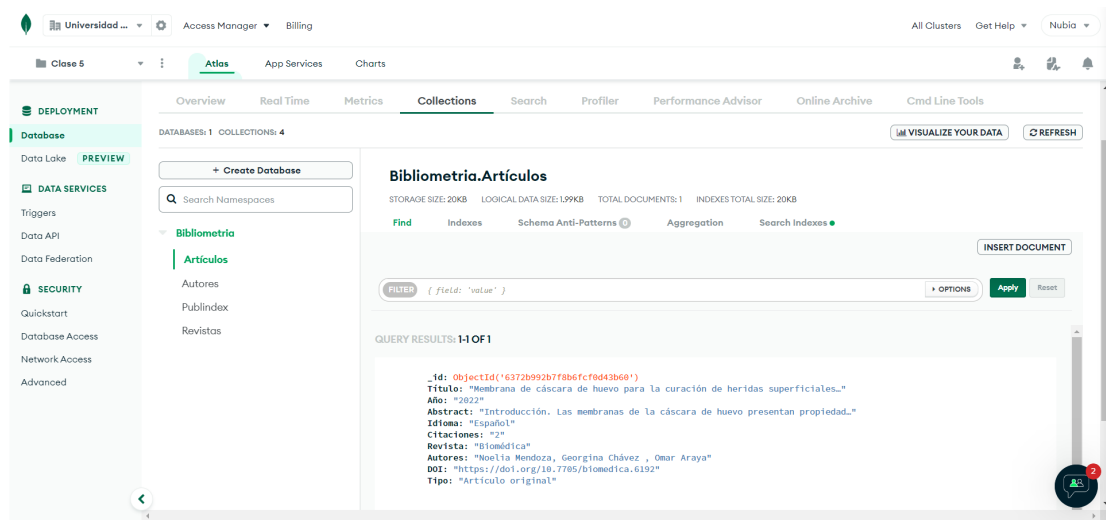


Figura 13: Base de Datos en MongoDB Atlas.

6. Bibliografía

Referencias

Martinovich, V. (2020). Indicadores de citación y relevancia científica: genealogía de una representación. *Dados*, 63(2), 1-27.

- Molina-Molina, S., Álvarez-Argaez, S., Estrada-Hernández, J., & Estrada-Hernández, M. (2020). Indicadores de ciencia, tecnología e innovación: hacia la configuración de un sistema de medición. *Revista Interamericana de Bibliotecología*, 43(3), 1-21.
- Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, 133, 285-296.
- Onodera, N., & Yoshikane, F. (2015). Factors affecting citation rates of research articles. *Journal of the Association for Information Science and Technology*, 66(4), 739-764.
- Repiso, R., Moreno-Delgado, A., & Aguaded, I. (2021). Factors affecting the frequency of citation of an article. *Iberoamerican Journal of Science Measurement and Communication*, 1(1), 007-007.
- Navarrete, L., & Pérez, C. (2019). Revistas Biomédicas: desarrollo y evolución. *Revista Médica Clínica Las Condes*, 30(3), 219-225.
- Anderson, C., Nugent, K., & Peterson, C. (2021). Academic journal retractions and the COVID-19 pandemic. *Journal of Primary Care & Community Health*, 12, 1-6.
- University of Michigan Library. (2022). *Research Impact Metrics: Citation Analysis - Scopus* [<https://guides.lib.umich.edu/citation/Scopus/> Recuperado el 04/10/2022].
- Elsevier. (2022). *Scopus content - How Scopus works* [https://www.elsevier.com/solutions/scopus/how-scopus-works/content?dgcid=RN_AGCM_Sourced_300005030/ Recuperado el 04/10/2022].
- Suehring, S. Getting started. En: En *MySQL Bible*. New York: Wiley Publishing, Inc., 2002. Cap. 1.
- Oracle. (2022). What is MySQL? <https://dev.mysql.com/doc/refman/8.0/en/what-is-mysql.html>
- Hernández Chillón, A., Morales, S. F., García Molina, J., & Sevilla Ruiz, D. (2017). *Visualización de Esquemas en Bases de Datos NoSQL basadas en documentos* [https://biblioteca.sistedes.es/submissions/uploaded-files/JISBD_2017_paper_71.pdf Recuperado el 14/11/2022].