

# Proyecto de Bases de datos para un análisis bibliométrico

Nubia Fernanda Sánchez Bello<sup>1</sup>

<sup>1</sup>Facultad de Ingeniería y Ciencias Básicas

Universidad Central

Maestría en Analítica de Datos

Curso de Bases de Datos

Bogotá, Colombia

<sup>1</sup>nsanchezb1@ucentral.edu.co

8 de octubre de 2022

## Índice

<b>1. Introducción (Max 250 Palabras) - (<i>Primera entrega</i>)</b>	<b>3</b>
<b>2. Características del proyecto de investigación (<i>Primera entrega</i>)</b>	<b>3</b>
2.1. Título del proyecto de investigación ( <i>Primera entrega</i> ) . . . . .	3
2.2. Objetivo general ( <i>Primera entrega</i> ) . . . . .	3
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> ) . . . . .	4
2.5. Hipótesis ( <i>Primera entrega</i> ) . . . . .	4
<b>3. Reflexiones sobre el origen de datos e información (<i>Primera entrega</i>)</b>	<b>5</b>
3.1. ¿Cual es el origen de los datos e información ? ( <i>Primera entrega</i> ) .	5
3.2. ¿Cuales son las consideraciones legales o eticas del uso de la información? ( <i>Primera entrega</i> ) . . . . .	5
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> ) . . . . .	6
3.4. ¿Que espera de la utilización de un sistema de Bases de Datos para su proyecto? ( <i>Primera entrega</i> ) . . . . .	6

<b>4. Diseño del Modelo de Datos del SMBD (Sistema Manejador de Bases de Datos)(Primera entrega)</b>	<b>7</b>
4.1. Características del SMBD (Sistema Manejador de Bases de Datos) para el proyecto ( <i>Primera entrega</i> ) . . . . .	7
4.2. Diagrama modelo de datos ( <i>Primera entrega</i> ) . . . . .	7
4.3. Imágenes de la Base de Datos ( <i>Primera entrega</i> ) . . . . .	8
4.4. Código SQL - lenguaje de definición de datos (DDL) ( <i>Primera entrega</i> ) . . . . .	9
4.5. Código SQL - Manipulación de datos (DML) ( <i>Primera entrega</i> ) . .	10
4.6. Código SQL + Resultados: Vistas ( <i>Primera entrega</i> ) . . . . .	13
4.7. Código SQL + Resultados: Triggers ( <i>Primera entrega</i> ) . . . . .	15
4.8. Código SQL + Resultados: Funciones ( <i>Primera entrega</i> ) . . . . .	16
4.9. Código SQL + Resultados: procedimientos almacenados ( <i>Primera entrega</i> ) . . . . .	17
<b>5. Bibliografía</b>	<b>18</b>

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

La bibliometría es un campo que existe desde los años 60 y consiste en la aplicación de métodos estadísticos para cuantificar procesos de comunicación escrita y el desarrollo de disciplinas científicas; la bibliometría tiene un importante uso como herramienta que evalúa algunos aspectos del desarrollo de campos o disciplinas científicas (Buitrago-Pulido, 2019).

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.

Las técnicas empleadas para los análisis bibliométricos son diversas, suelen enfocarse en dos aspectos principales: el análisis de desempeño, y el mapeo de la ciencia, siendo el primero lo referente a revisar las contribuciones de los participantes de una investigación a un área determinada (número de publicaciones, número de citaciones, proporción de publicaciones citadas), y el segundo, enfocado a establecer las relaciones que se presentan al realizar un proceso de investigación (influencia de las publicaciones, co-autoría, relaciones entre temas). Los alcances que pueden tener estos aspectos principales se han visto mejorados por técnicas complementarias como métricas de red, visualización en software especializado y análisis a través de clustering (Donthu et al., 2021).

El avance en la bibliometría con las técnicas de análisis mejorado y la obtención de cantidades masivas de información actualizada constantemente, permite identificar tendencias y vacíos de conocimiento, así como investigadores e instituciones clave dentro de una disciplina. Estos análisis pueden emplearse como estrategia para reconocer hacia donde avanza el conocimiento de un área, qué aspectos pueden mejorarse y cuáles actores clave pueden aportar estos procesos. Resulta entonces de interés enfocar este tipo de análisis en áreas de crecimiento constante, como por ejemplo la literatura científica biomédica.

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

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

Análisis bibliométrico de la producción científica de revistas científicas biomédicas colombianas indexadas en Scopus desde 2004.

### **2.2. Objetivo general (*Primera entrega*)**

Realizar un análisis bibliométrico de los artículos publicados a partir de 2004 en las revistas científicas biomédicas colombianas indexadas en Publindex y en Scopus.

### **2.2.1. Objetivos específicos (*Primera entrega*)**

- Identificar tendencias de los artículos publicados por las revistas científicas biomédicas en el período de análisis.
- Examinar el comportamiento bibliométrico de cada revista científica biomédica durante el período de análisis y compararlo.
- Categorizar las principales temáticas presentadas por las revistas científicas biomédicas.
- Identificar a los principales actores que generaron la producción científica de las revistas científicas biomédicas colombianas en el período de análisis.

### **2.3. Alcance (*Primera entrega*)**

El alcance de este proyecto es, en principio, descriptivo. 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, por su naturaleza, diversidad y contexto, no se espera que sea posible realizar asociaciones entre variables que permitan diferenciar con claridad una relación.

### **2.4. Pregunta de investigación (*Primera entrega*)**

¿Cuál ha sido el comportamiento de la producción científica de las revistas biomédicas colombianas a partir del año 2004?

### **2.5. Hipótesis (*Primera entrega*)**

La producción científica de las revistas científicas biomédicas colombianas se ha ido incrementando a lo largo del tiempo, sus temáticas se han vuelto más diversas y se han generado redes de trabajo interdisciplinar para fomentar su desarrollo.

### **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)

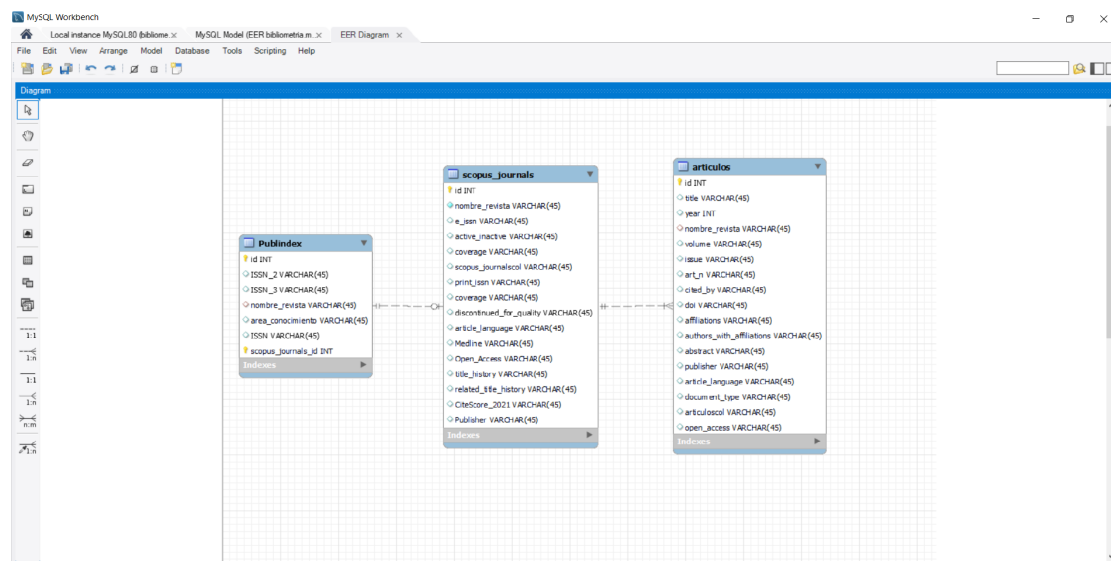


Figura 1: Primera versión del diagrama modelo de datos.

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

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

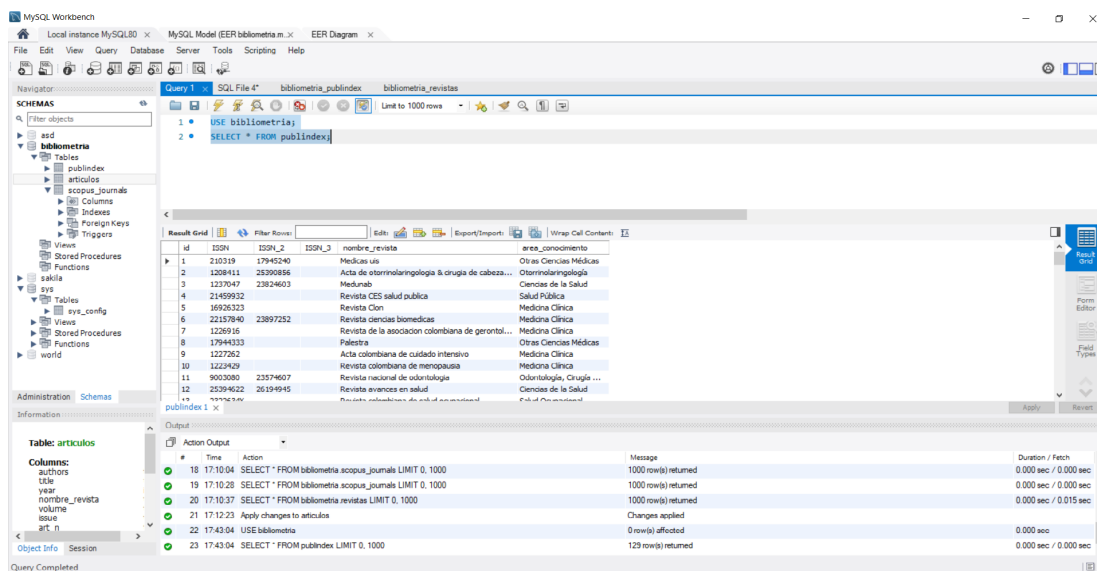


Figura 2: Imagen tabla Publindex.



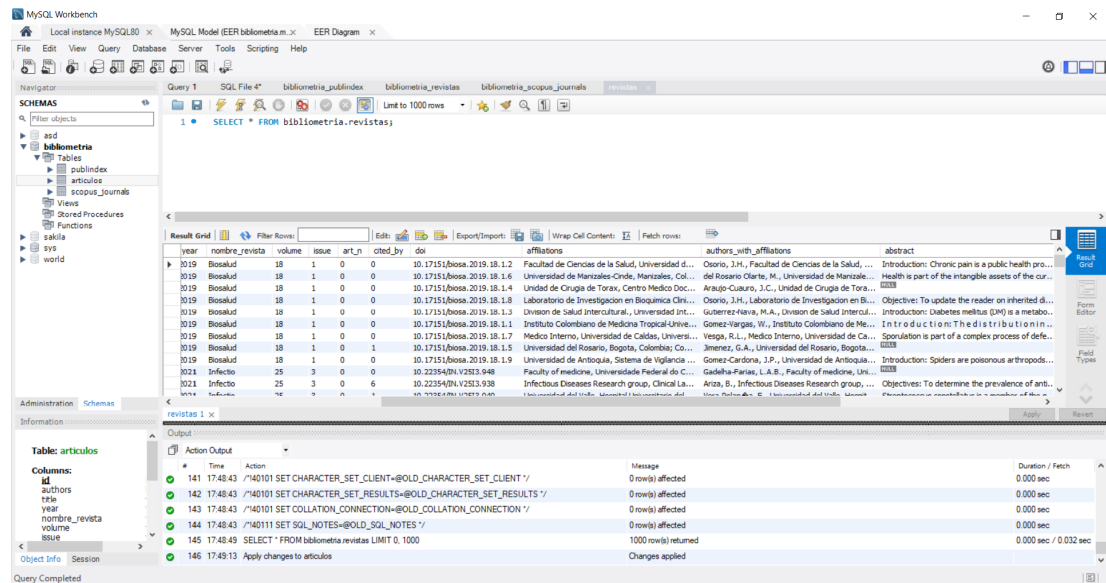


Figura 3: Imagen tabla Artículos.

#### 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.

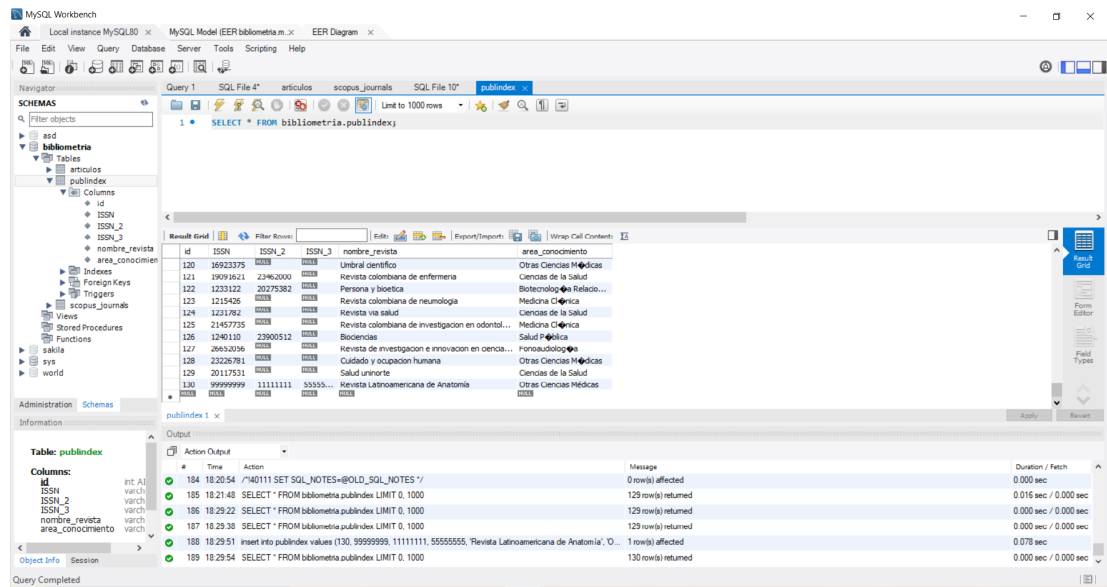


Figura 4: Adición de nuevo registro a tabla Pubindex.

Ahora con update se actualiza el ISSN de este registro:

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

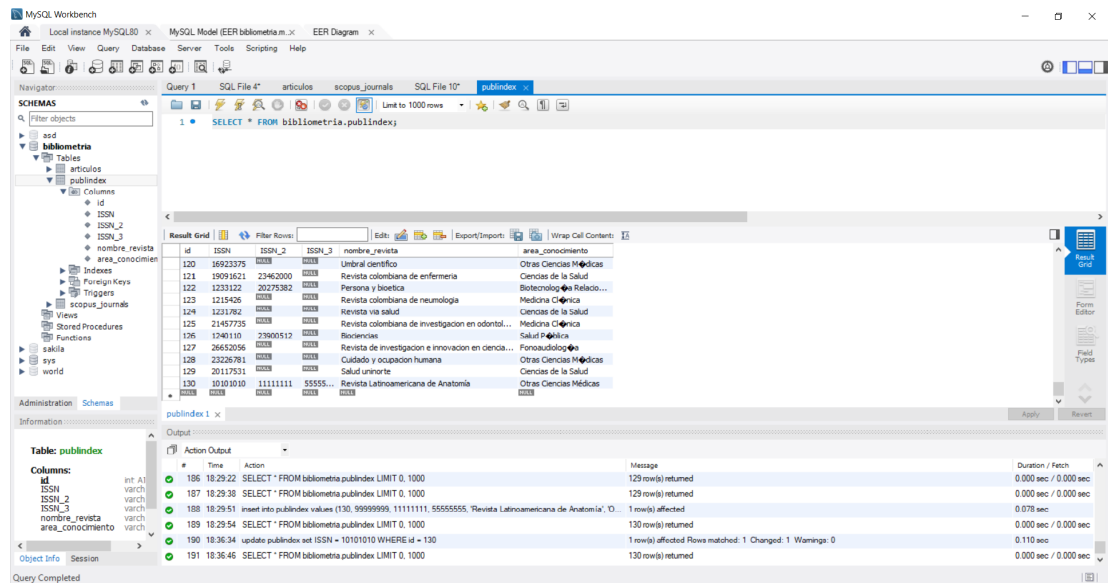


Figura 5: Actualización de registro en tabla Publindex.

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;
```

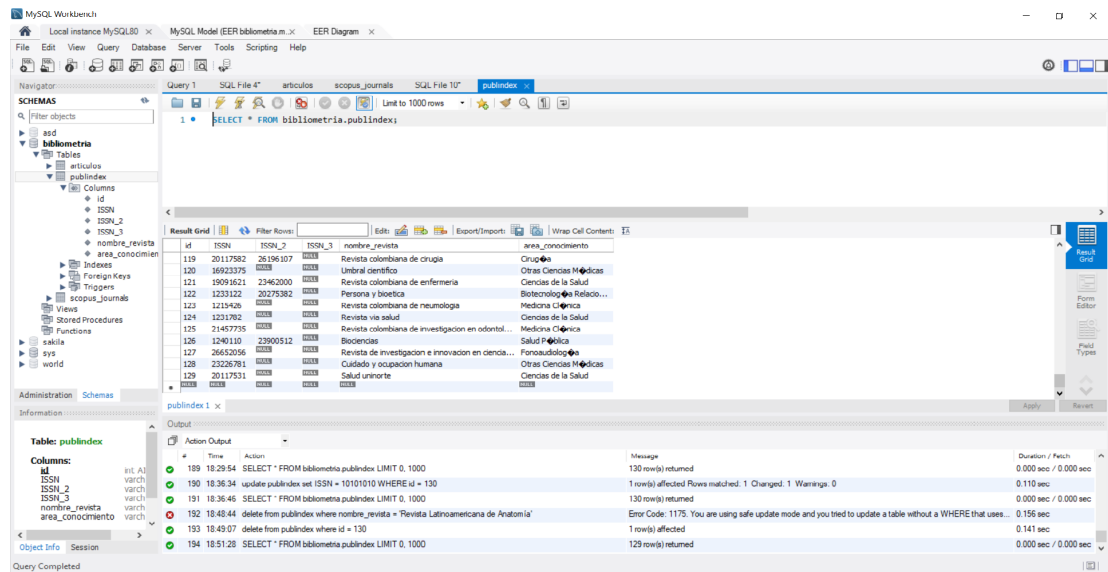


Figura 6: Registro eliminado de la tabla Publindex.

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

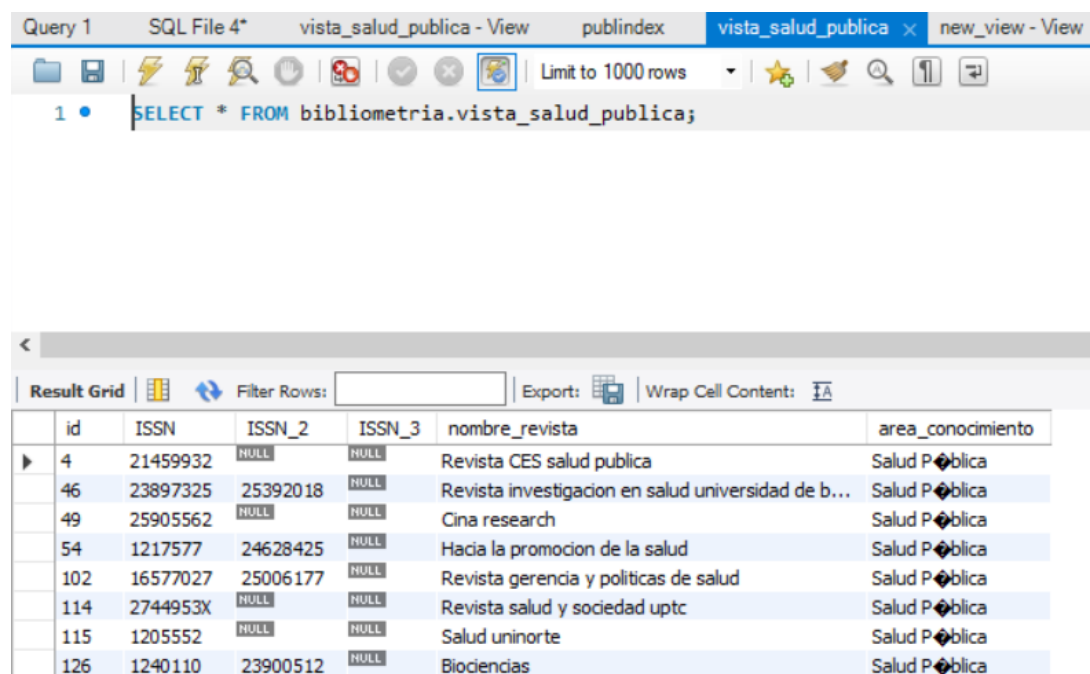
Se presentan algunas vistas de relevancia para análisis de información. En primer lugar una vista que me permita ver revistas de un área en específico, por ejemplo, las revistas colombianas de Salud Pública:

```

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    publindex    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: **TA**

	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:

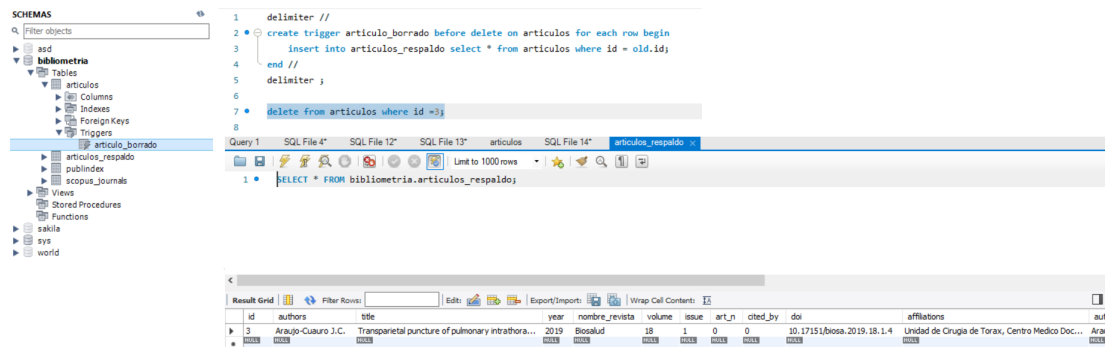


Figura 9: Resultados del trigger para almacenar artículos borrados.

#### 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:

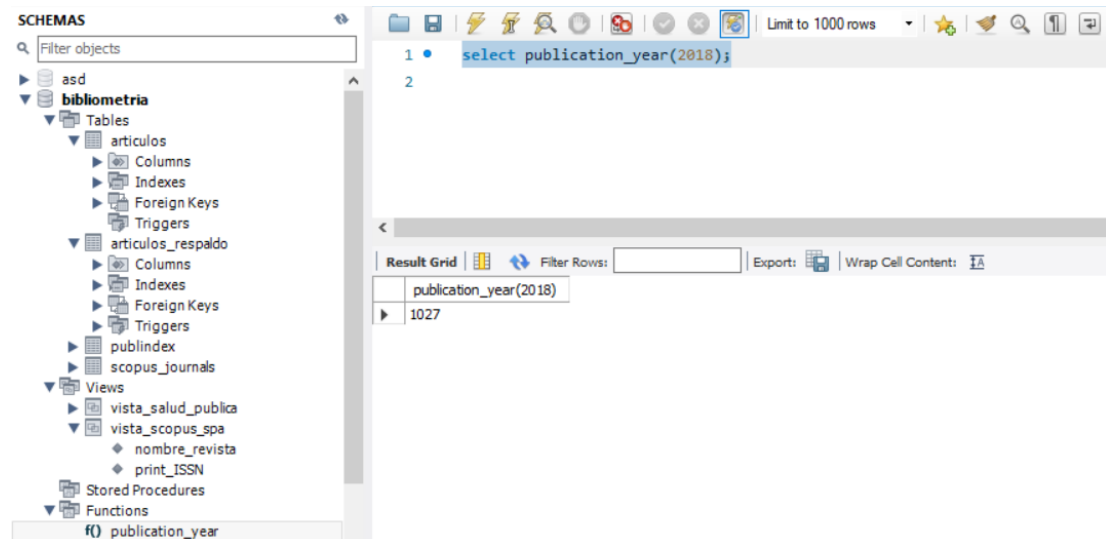


Figura 10: Número de artículos publicados en el año 2018.

#### 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 MySQL database interface. On the left, the 'SCHEMAS' panel lists databases: 'asd', 'bibliometria', 'sakila', 'sys', and 'world'. The 'bibliometria' database is selected, and the 'indexados' table is highlighted. The main window displays the 'indexados' table with two columns: 'nombre\_revista' and 'issn'. The table contains 15 rows of data, including journals like 'Aquichan', 'Archivos de medicina', 'Biomedica', 'Biosalud', 'Colombia medica', 'Hacia la promocion de la salud', 'Iatreia', 'Infectio', 'Investigacion y educacion en enfermeria', 'Medicina', and 'Medicina'.

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
Medicina	1205498
Medicina	1205498
Medicina	1205498

Figura 11: Revistas indexadas en Publiindex y en Scopus.

## 5. Bibliografía

### Referencias

- Buitrago-Pulido, R. D. (2019). Análisis bibliométrico sobre la producción científica en distribución en planta en la red Redalyc durante el periodo 2007-2017. *Scientia et technica*, 24(3), 446-450.
- 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.
- 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: *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