# DESCRIPCIÓN DE LA BASE DE DATOS DE PUBLICACIONES CIENTÍFICAS SEMANTIC SCHOLAR

## 1. Control de Versiones

Versión actual: 1.0

Fecha: julio de 2019

Autor: Jerónimo Arenas García

## 2. Objetivo del documento

El presente documento describe la base de datos MySQL obtenida a partir de los datos abiertos publicados por Semantic Scholar.

<https://api.semanticscholar.org/corpus/download/>

La presente versión está basada en los datos publicados por Semantic Scholar con fecha de enero de 2019.

## 3. Descripción de la base de datos

La base de datos consta de cinco tablas:

* S2papers: 46.947.044 artículos.
* S2authors: 17.805.885 autores.
* S2PaperAuthor: 156.371.021 pares (autor-paper), dando una media de 3,33 autorías por artículo incluido en la base de datos.
* citations: 348.041.657 de citas, dando una media de 7,41 citas por artículo.
* S2journals: 53.246 revistas únicas en la base de datos.
* S2venues: 79.131 venues únicos en la base de datos.

Analizamos a continuación la estructura de la base de datos.

#### 3.1. Tabla S2papers

Contiene fundamentalmente la información relativa a cada artículo científico, que ha sido en su mayoría extraída de la propia información volcada por Semantic Scholar. Además, se han añadido algunos campos adicionales de interés para las actividades del Proyecto Faro:

A close up of text on a black background

Description automatically generated

##### Campos obtenidos de Semantic Scholar

* paperID: Un identificador inherente a la propia base de datos. Se ha decidido crear este índice adicional para acelerar la actualización índice primario durante la carga de los datos.
* S2paperID: El identificador propio de Semantic Scholar para cada paper. Se trata de una cadena alfanumérica de 40 caracteres. En principio, los scripts de importación no indexan el contenido de este campo. No obstante, si se van a realizar búsquedas empleando el valor de este campo como criterio, puede ser una buena idea disponer de dicho índice:

mysql> CREATE INDEX S2paperID on S2papers(S2paperID);

* S2PdfUrl: Cuando este campo tiene un valor no nulo, existe un PDF disponible a partir de la propia web de Semantic Scholar para el artículo en cuestión. En ese caso, podríamos acceder a dicho documento a partir de la URL disponible en este canmpo.
* pdfUrls: Distintas URLs que permitirían acceder al pdf del paper, aunque no todas ellas conducen realmente al artículo, sino a páginas que permitirían su descarga si se tiene adquirido el derecho. Cuando existen más de una direcciones listadas en este campo, éstas vienen separadas por un tabulador “\t”.
* year: Contains the year in which the paper was published. There are 874.619 papers without a valid publication year, that have been identified in the database using the year value ‘9999’. The number of papers per year is illustrated in the following figure



* venueID: A numeric value with the ID corresponding to the venue of the paper. The full name of the venue can be found in the S2venues table, which is a table that contains just this information for the 79.131 venues included in the dataset:

The table S2 venues contains just two fields: venueID and venue.


* journalNameID: A numeric value with the ID corresponding to the journal of the paper. The full name of the journal can be found in the S2journals table, which is a table that contains just this information for the 53.246 journals included in the dataset:

A black and red text

Description automatically generated

* journalVolume / journalPages: Two text fields containing information regarding the number, volume, and publication pages. It has been included in the database for completeness, but it is very unlikely we will use this information.
* isDBLP: Contains “1” for papers that are included in the DBLP repository, which basically aggregates papers in the field of computer science. There are 4.313.127 papers with this flag activated.
* isMedline: Contains “1” for papers that are included in the Medline repository (<https://es.wikipedia.org/wiki/MEDLINE>). There are 29.317.121 papers with this flag activated.

There are also 217.757 papers that have both flags (isDBLP and isMedline) activated.

* doi: Digital Object Identifier registered at [doi.org](http://www.doi.org/).
* doiUrl: DOI link for registered objects.
* pmid: Unique identifier used by PubMed.

##### Campos textuales

* title: Título de la publicación
* lowertitle: Versión en minúsculas del título del paper. Se genera esta versión en minúsculas para poder cruzar de forma más sencilla con otras bases de datos.
* paperAbstract: Resumen del paper.
* entities: Entidades identificadas por Semantic Scholar. Para cada paper, las entidades disponibles se incluyen separadas por un tabulador.

En total, se dispone de 481.012 entidades para todo el corpus, si bien no todas ellas aparecen en la misma medida. Así, la siguiente figura muestra que más de 100.000 entidades aparecen únicamente en un paper, mientras que el número de entidades que aparece en más de 1.000 artículos se sitúa en torno a las 40.000.



* LEMAS: LEMAS extraídos del título y resumen del artículo.

##### Generación de nuevos metadatos

* ESP\_contri:
* AIselection:

#### 3.2. Tabla citations

A green and black text

Description automatically generated

Incluye información sobre las citas detectadas por Semantic Scholar. Cada fila de la tabla corresponde a una cita en la que se indica que paperID1 ha citado a paperID2. Se incluye un campo adicional que de momento no se ha rellenado:

* isInfluential: En Semantic Scholar se realiza un análisis basado en grafos para determinar que papers de los citados por otro artículo son los que realmente lo han influenciado.

#### 3.3. Tabla S2authors

Contiene fundamentalmente la información relativa a autores volcada por Semantic Scholar: ID de autor y su nombre. Además, se han añadido algunos campos adicionales de interés para la desambiguación de autores o para otras actividades del Proyecto Faro:

A close up of text on a white background

Description automatically generated

##### Campos obtenidos de Semantic Scholar

* authorID: el identificador propio de Semantic Scholar para cada autor.
* name: El nombre principal incluido en Semantic Scholar para cada autor.

##### Campos que podrían obtenerse via API de Semantic Scholar

* influentialCitationCount: Es un conteo de cuantos artículos de este autor han influenciado a otros artículos, y por lo tanto podría emplearse como una medida de la relevancia del autor.

##### Generación de nuevos metadatos: cruce con base de datos de ORCID

El esquema de la base de datos incluye otros campos que pueden ser relevantes para la tabla de autorías. Se trataría fundamentalmente de: 1/ detectar qué autores están sólidamente identificados en S2; 2/ cruzar con otras bases de datos, y detectar qué autores han tenido en algún momento afiliación a una entidad española.

* ORCIDID, ORCID\_givename, ORCID\_familyname: Nombre y apellidos recogidos en la base de datos de ORCID
* SCOPUSID: Identificador de SCOPUS para el autor
* ESP\_affiliation:

#### 3.4. Tabla PaperAuthor

Se trata de una tabla que contiene una entrada por cada par paper-autor, donde la existencia de dicha entrada indica que el autor con authorID es (co)autor del paper con paperID.

A white sign with black text

Description automatically generated

Según se mencionó anteriormente, existen casi 18 millones de autores incluidos en la base de datos, pero el número de artículos asociados a cada autor varía de forma importante según se muestra en la siguiente figura. Así, podemos ver que el número de autores que ha publicado más de un artículo es inferior a los 12 millones, o que el número de autores con 10 o más publicaciones se sitúa alrededor de los 3 millones.

