**UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA**

****

**SISTEMAS INFORMÁTICOS Y COMPUTACIÓN**

**Sistemas basados en Conocimiento**

**Nombre:**

**Anthony Ortiz**

**DOCENTE:**

**Janneth Alexandra Chicaiza**

**Fecha:**

**05-08-2020**

**Loja - Ecuador**

Contenido

[Introducción 3](#_Toc47476192)

[Objetivos 4](#_Toc47476193)

[Contenidos 4](#_Toc47476194)

[Especificación de la ontología 4](#_Toc47476195)

[Construcción de glosario de términos 5](#_Toc47476196)

[Construir taxonomía de conceptos 6](#_Toc47476197)

[Construir diagrama de relaciones binarias 7](#_Toc47476198)

[Transformación de datos RDF 8](#_Toc47476199)

[Consultas SPARQL 14](#_Toc47476200)

[Aplicación propuesta 17](#_Toc47476201)

[Conclusiones 18](#_Toc47476202)

[Bibliografía 19](#_Toc47476203)

# Introducción

Estudios conservadores recientes estiman que los sistemas de servidores de las empresas en el mundo han procesado 9,57×10 bytes de datos en 2008. Se espera que este número se haya duplicado cada dos años a partir de ese momento. Como ejemplo, los servidores de Walmart manejan más de un millón de transacciones de clientes cada hora, y esta información se inserta en bases de datos que almacenan más de 2,5 petabytes de datos, el equivalente a 167 veces el número de libros de la Biblioteca del Congreso.

Estos ejemplos ofrecen un pequeño vistazo al ecosistema en rápida expansión de diversas fuentes de conjuntos de datos masivos que existen actualmente. Los datos pueden ser estructurados (por ejemplo, financieros, registros médicos electrónicos, estadísticas gubernamentales), semiestructurados (por ejemplo, texto, tweets, correos electrónicos), no estructurados (por ejemplo, audio y vídeo) y en tiempo real (por ejemplo, trazas de redes, registros genéricos de vigilancia). Todas estas aplicaciones comparten el potencial de proporcionar conocimientos inestimables, si se organizan y analizan adecuadamente (Kambatla, K., Kollias, G., Kumar, V., Grama; 2014).

Sin controlar los datos generados a cada minuto es un desperdicio por ello investigadores han creado formas de representar los datos en información de manera estructurada y entendible, mediante el uso de Tripletas-RDF. Esta base deconocimiento consiste típicamente en triples marcos de descripción de recursos (RDF) que describen las entidades y sus relaciones. La generación de la descripción en lenguaje natural de la base de conocimientos es una tarea importante en el PNL, que se ha formulado como una tarea de generación de lenguaje condicional y se ha abordado utilizando el marco de secuencia a secuencia (Zhu, Y., Wan, J., Zhou, Z., Chen, L., Qiu, L., Zhang, W., Yu, Y; 2019).

En este documento, se creará una base de conocimientos en forma de triples RDF, el objetivo es generar una descripción en lenguaje natural de publicaciones provenientes de investigaciones Covid-19 que sean gramaticalmente correctas, fáciles de entender y capaces de entregar la información mediante el uso de una aplicación. El repositorio utsado para almacenar los archivos de todo el proyecto es:

https://github.com/AnthonyOrtiz/Proyecto\_SBC\_2020.git

# Objetivos

Objetivo general:

* Desarrollar una aplicación que permita visualizar e interactuar con información de datos RDF generados a partir de un conjunto de dataset recolectada en relación con las investigaciones sobre COVID-19 o SARS-COV-2.

Objetivos específicos:

* Crear un modelo ontológico relacionada a publicaciones COVID-19.
* Extraer metadatos relacionados a publicaciones COVID-19.
* Generar datos RDF con base a un modelo ontológico sobre publicaciones COVID-19.
* Realizar consultas SparQL hacia un repositorio semántico con la finalidad de visualizar los datos a través de gráficas y tablas estadísticas.

# Contenidos

En este documento, se creará una base de conocimientos en forma de triples RDF, el objetivo es generar una descripción en lenguaje natural de publicaciones provenientes de investigaciones Covid-19 que sean gramaticalmente correctas, fáciles de entender y capaces de entregar la información mediante el uso de una aplicación.

## Especificación de la ontología

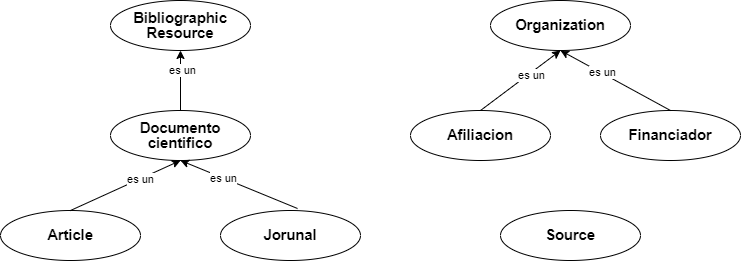
En esta sección se describe la especificación de la ontología que se usará para los datos RDF de publicaciones COVID-19.

|  |  |
| --- | --- |
| SEEMP Reference Ontology Requirements Specification Document | |
| **1** | **Purpose** |
|  | El propósito de esta ontología es proveer de un modelo de conocimiento de artículos o documentos sobre COVID 19 que será implementada como información semántica. |
| **2** | **Scope** |
|  | La ontología solo se centra en las investigaciones científicas por palabras claves sobre el COVID 19. |
| **3** | **Implementation Language** |
|  | La ontología tiene que ser implementada en el lenguaje OWL. |
| **4** | **Intended End-Users** |
|  | User 1. Investigadores y personas particulares que buscan comprender sobre los aportes científicos del COVID 19.  User 2. Información y datos de los documentos científicos sobre el COVID 19. |
| **5** | **Intended Uses** |
|  | Use 1. Almacenar datos y proporcionar información sobre la diversidad y la composición de los artículos científicos.  Use 2. Proponer una norma y una práctica de gestión de los datos disponibles sobre documentos de COVID 19. |
| **6** | **Ontology Requirements** |
|  | a. Non-Functional Requirements |
|  | NFR1: La ontología debe soportar un escenario multilingüe en los siguientes idiomas: Inglés, español, italiano y francés.  NFR2: La ontología debe basarse en normas internacionales, europeas y de hecho en existencia o en desarrollo. |
|  | b. Functional Requirements: Groups of Competency Questions |
|  | 1. ¿Qué palabras claves y cuantas existen en los artículos? 2. ¿Cuáles son las palabras claves que más se comunes entre documentos? 3. ¿Cuál es la plataforma (databse) que contiene más artículos científicos sobre el COVID 19? 4. ¿Número de publicaciones por tipo de documento y numero de publicaciones por área de investigación? 5. ¿Cuál es el número de autores y cuantos son? 6. ¿Cuáles son las palabras indexadas a las palabras claves? 7. ¿Cuál es el número de publicaciones y el idioma al que pertenece? 8. ¿Cuáles son los nombres de los países con mayores publicaciones de COVID19? 9. ¿Cuáles son las publicaciones que tiene más financiadores? 10. ¿Cuáles son las organizaciones que han financiado mas los proyectos? |

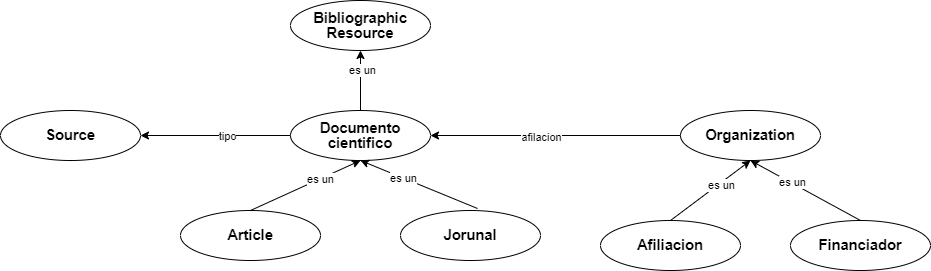
## Construcción de glosario de términos

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Nombre | Sinónimos | Acrónimos | Descripción | Tipo |
| **Bibliographic Resource** |  |  | Artículo científico que describe investigaciones encontradas por la pandemia CODIV 19 | Concepto |
| creator |  |  | Autores que contribuyeron con el articulo | Atributo de instancia |
| title |  |  | Título del articulo | Atributo de instancia |
| date |  |  | Fecha de publicación del documento | Atributo de instancia |
| type |  |  | Tipo de articulo al que pertenece el paper | Atributo de instancia |
| volume |  |  | Volumen que tiene el recurso | Atributo de instancia |
| issue |  |  | Issue que tiene un el recurso | Atributo de instancia |
| Citey By |  |  | Veces que ha sido citado | Atributo de instancia |
| identifier |  |  | Identificador del paper | Atributo de instancia |
| url |  |  | Doi del recurso bibliográfico | Atributo de instancia |
| comment |  |  | Abstract del articulo | Atributo de instancia |
| keywords |  |  | Palabras claves que contiene el paper | Atributo de instancia |
| publisher |  |  | La plataforma donde fue publicado | Atributo de instancia |
| language |  |  | Lenguaje sobre en el que fue escrito | Atributo de instancia |
| **Source** |  |  | Recurso detallado sobre el tipo de documento del recuro bibliográfico | Concepto |
| type |  |  | Tipo de articulo al que pertenece | Atributo de instancia |
| name |  |  | Nombre del recurso | Atributo de instancia |
| **Organization** |  |  | Organización de la cual fue financiada el recurso bibliográfico | Concepto |
| subject |  |  | Categorías de dbpedia a la que pertenece | Atributo de instancia |
| type |  |  | Tipo de organización que corresponde | Atributo de instancia |
| location |  |  | Lugar donde se creó el recurso bibliográfico | Atributo de instancia |

## Construir taxonomía de conceptos



## Construir diagrama de relaciones binarias



Describir las relaciones binarias ad hoc

|  |  |  |  |
| --- | --- | --- | --- |
| Nombre de la relación | Concepto de origen-dominio | Cardinalidad máxima | Concepto destino-rango |
| tipo | Documento | 1 | Source |
| afiliación | Documento | N | Organization |

Describir atributos de instancia

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Nombre del atributo | Concepto | Tipo valor | Rango de valores | Cardinalidad |
| creator | **Bibliographic Resource** | Cadena de caracteres |  | 1,N |
| title | **Bibliographic Resource** | Cadena de caracteres |  | 1,1 |
| date | **Bibliographic Resource** | Cadena de caracteres |  | 1,1 |
| type | **Bibliographic Resource** | Cadena de caracteres |  | 1,1 |
| volume | **Bibliographic Resource** | Cadena de caracteres |  | 1,1 |
| issue | **Bibliographic Resource** | Cadena de caracteres |  | 1,1 |
| Citey By | **Bibliographic Resource** | Cadena de caracteres |  | 1,1 |
| identifier | **Bibliographic Resource** | Cadena de caracteres |  | 1,1 |
| url | **Bibliographic Resource** | Cadena de caracteres |  | 1,1 |
| comment | **Bibliographic Resource** | Cadena de caracteres |  | 1,1 |
| keywords | **Bibliographic Resource** | Cadena de caracteres |  | 1,N |
| publisher | **Bibliographic Resource** | Cadena de caracteres |  | 1,1 |
| language | **Bibliographic Resource** | Cadena de caracteres |  | 1,1 |
| type | **Source** | Cadena de caracteres |  | 1,1 |
| name | **Source** | Cadena de caracteres |  | 1,1 |
| subject | **Organization** | Cadena de caracteres |  | 1,N |
| type | **Organization** | Cadena de caracteres |  | 1,N |
| location | **Organization** | Cadena de caracteres |  | 1,N |

# Transformación de datos RDF

En esta sección se describe los pasos necesarios desde la extracción de datos sobre publicaciones COVID-19 hasta la generación de datos RDF con Jena, un framework para la construcción de ontologías. Extracción de datos

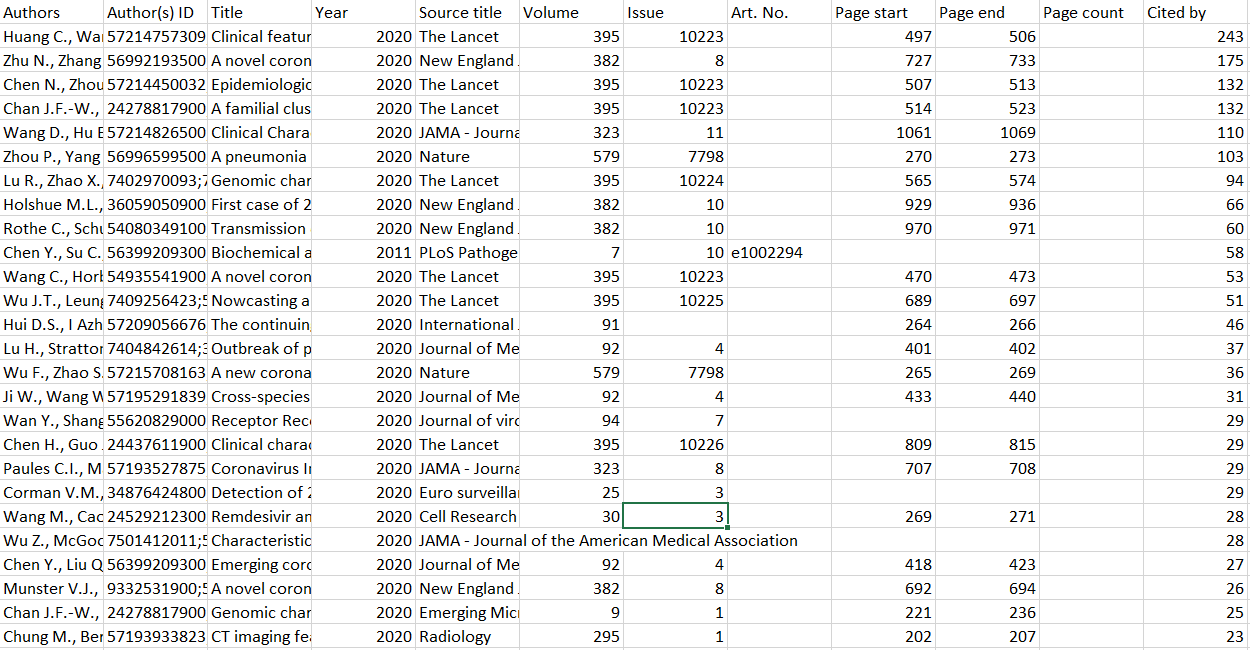
La fuente de datos en donde se extrajo las publicaciones científicas sobre Covid-19 fue en la plataforma SCOPUS, una base de datos científica en donde se puede encontró publicaciones en diferentes categorías.

Limpieza de datos

Con los datos disponibles, el siguiente paso es refinarlo para evitar problemas de caracteres con las generaciones de las tripletas RDF que puedan provocar. Para la limpieza de datos se realizó las siguientes actividades:

* Transformar el formato del archivo CSV a un formato legitimo para que pueda ser compatible con el formato de la versión de Windows 10.
* Agrupar campos en una sola celda
* Eliminar caracteres especiales de cada registro

El resultado final es un archivo mucho mas preparado y estructurado listo para realizar la extracción de los datos pulidos para la creación de las tripletas el resultado se puede apreciar en la siguiente imagen.



Selección de URIS

Para el uso de los prefijos para la generación de las tripletas se basaron en las siguientes:

|  |  |  |
| --- | --- | --- |
| **Prefijo** | **URI** | **Tipo** |
| **data** | http://utpl.edu.ec/lod/publicCOVID/data/ | Definida |
| **myOnto** | http://utpl.edu.ec/lod/publicCOVID/myOnto/ | Definida |
| **dc** | http://purl.org/dc/terms/ | Establecida en la web |
| **schema** | http://schema.org/ | Establecida en la web |
| **fabio** | http://purl.org/spar/fabio/ | Establecida en la web |
| **rdfs** | https://www.w3.org/2000/01/rdf-schema/ | Establecida en la web |
| **bibo** | http://purl.org/ontology/bibo/ | Establecida en la web |
| **dbc** | http://dbpedia.org/resource/ | Establecida en la web |

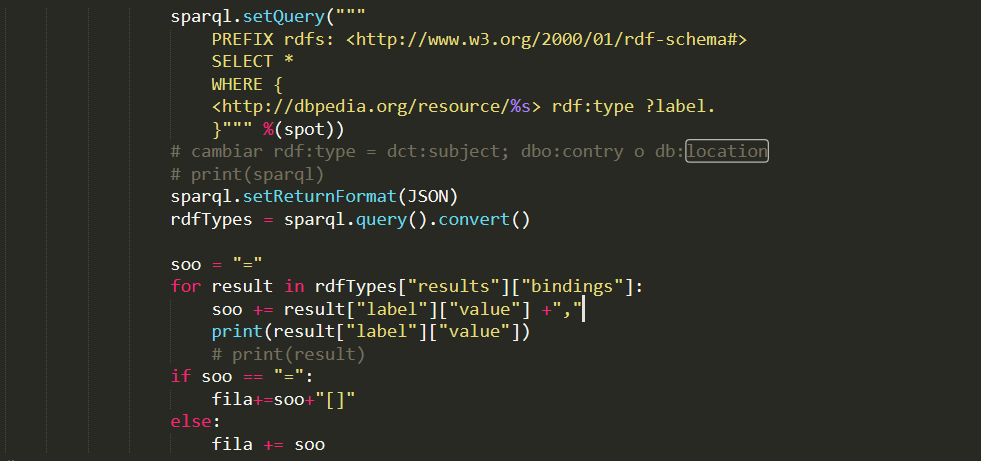
Generación Jena

En esta sección se explicará la creación de las Tripletas-RDF que se toma como base los datos del archivo CSV con la finalidad de generar un nuevo archivo con extensión rdf. Este proceso se lo realizo mediante el lenguaje de programación Java con el IDE “Apache Netbeans IDE 12” y la librería Jena que proporciona todas las funcionalidades necesarias. Además, se realizó en enriquecimiento de los datos del archivo CSV mediante el lenguaje Python gracias a la API de “tagme” que permite la información de nombre de organizaciones mediante texto.

El primer paso fue realizar consultas a la Api de tagme para obtener algunos datos de las publicaciones o investigaciones. Este proceso tiene como resultado buscar si un paper o investigación es afiliado a una organización, el texto se lo obtuvo del mismo archivo csv donde se encuentra las publicaciones y este se manda a tagme donde nos retornara las respuesta mediante un REQUEST (Ver la siguiente figura).



Al mismo tiempo que se hace esta consulta se realiza otra por medio de un ENDPOINT a dbPedia para obtener datos de las organizaciones que obtuvimos con la anterior consulta, esto se lo puede apreciar en la siguiente imagen.



Esta nueva información será generada en un arcihvo cv, para después extraer los datos y acoplarlos al archivo original de las investigaciones.

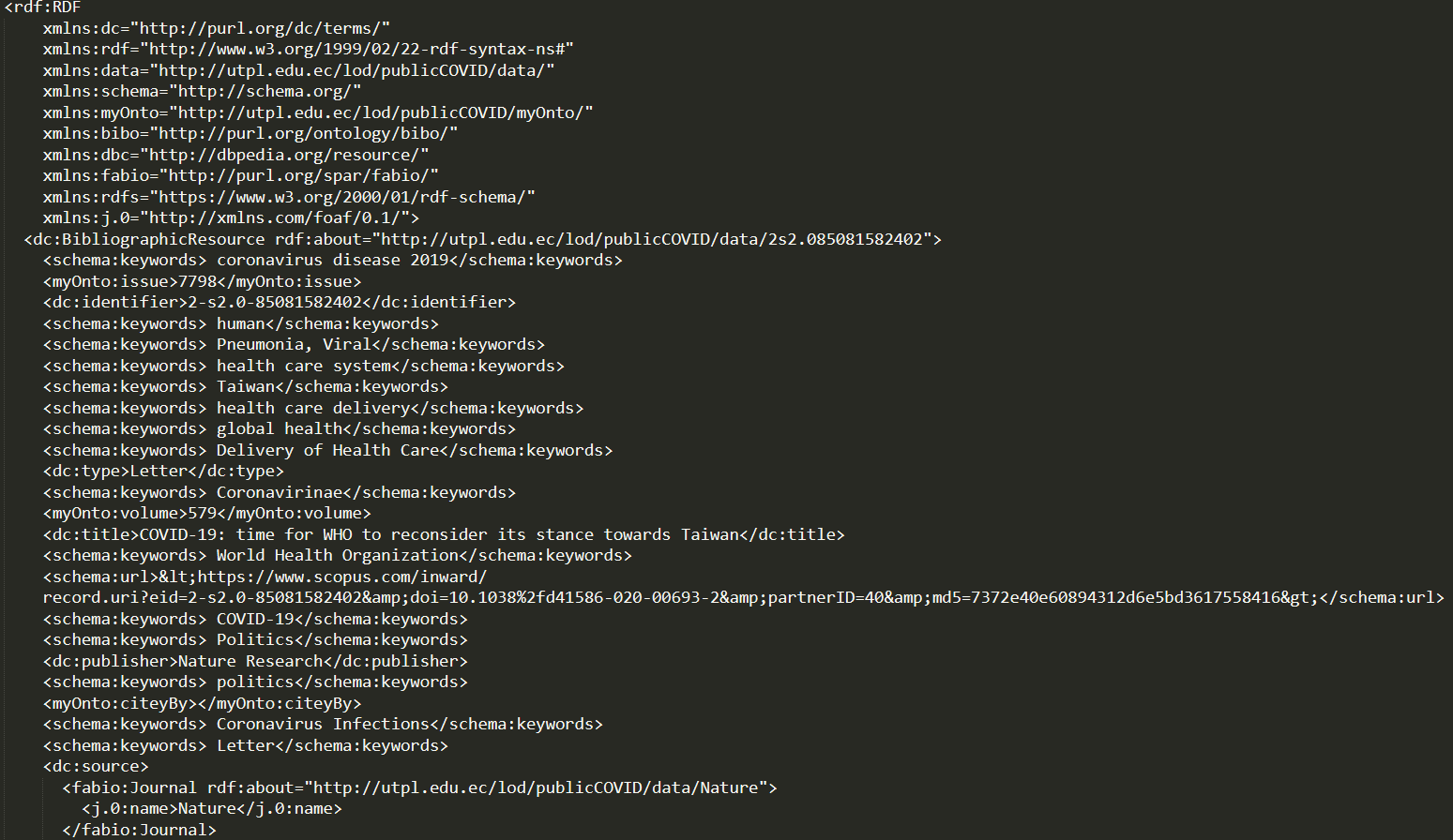
Una vez obtenido todos los datos necesarios se procedio a generar el archivo rdf con la librería Jena. El proceso para la generación de datos RDF consiste en primer lugar establecer los prefijos a un modelo, ver la siguiente figura.



Después se realizo una lectura de los datos del archivo csv donde se encuentran las publicaciones y esta lectura almacena los datos en un ArrayList, estos son leídos y almacenados en clases creadas para tener un mejor control de los campos de las publicaciones al momento de crear el rdf. Estos datos son obtenidos por medio de una estructura “for” para generar las triples que se pueden apreciar en la siguiente imagen.

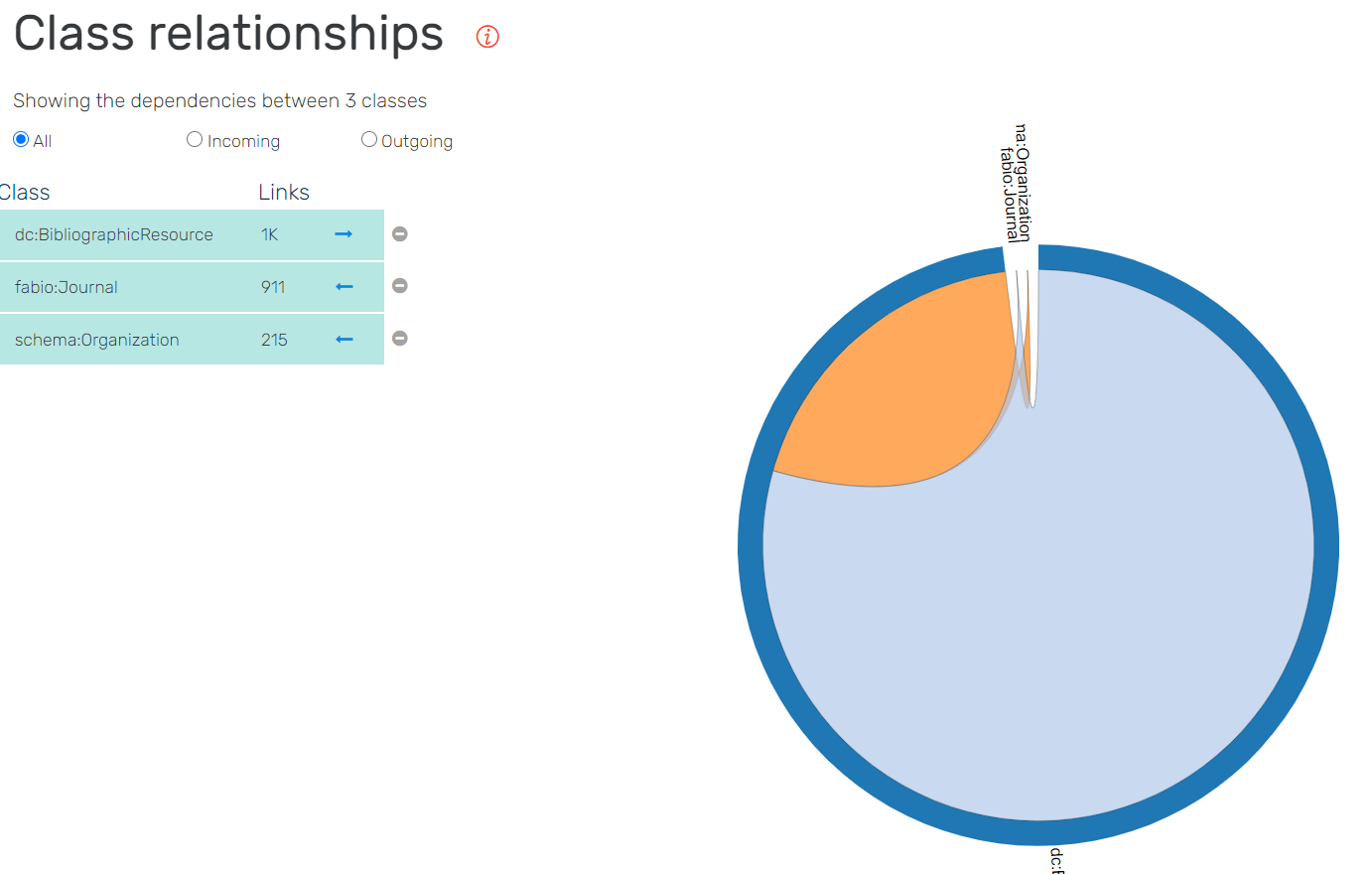


En la siguiente figura se puede apreciar un extracto sobre las tripletas generadas por Jena.



Almacenamiento de los datos RDF

Antes de poder usar estos datos estructurados en alguna aplicación es necesario subirlos en una base de datos para tener un control mucho mas gestionado sobre la información de las publicaciones, por ello se uso la base de grafos denominada “GraphDB”. En la siguiente imagen se puede apreciar los datos subidos a la base de datos.

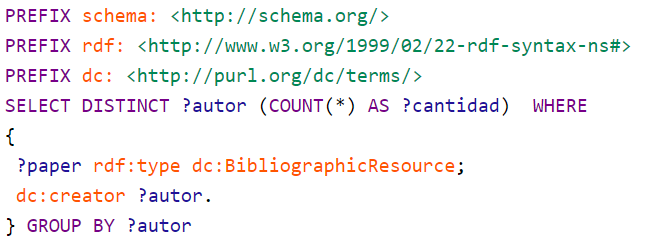


# Consultas SPARQL

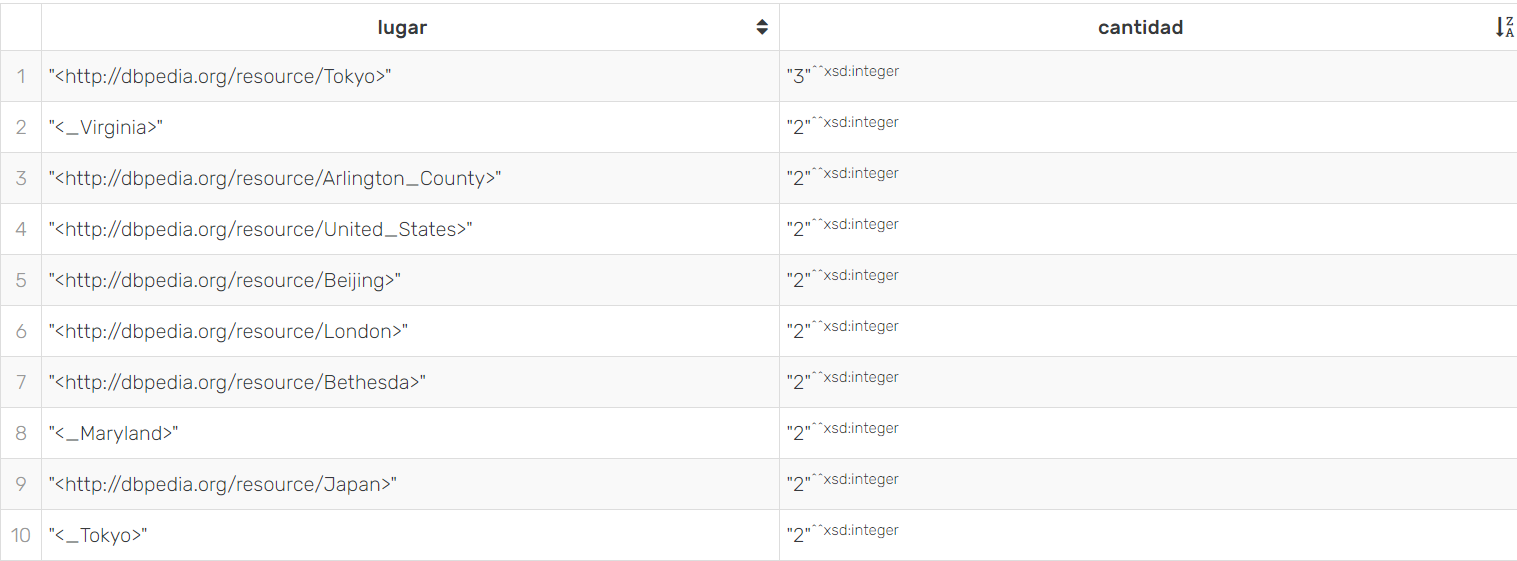
En Se ha creado apis en el servidor en donde cada una genera consultas hacia el repositorio semántico. Y con la información que se obtiene se procede a generar gráficas y tablas estadísticas en las interfaces de usuario. Las consultas que se han generado a partir de los datos que se encuentra en GrahpDB fueron las siguientes:

**Pregunta:** Consulta sobre la cantidad las participaciones de un autor en las publicaciones de covid-19

**Consulta**:

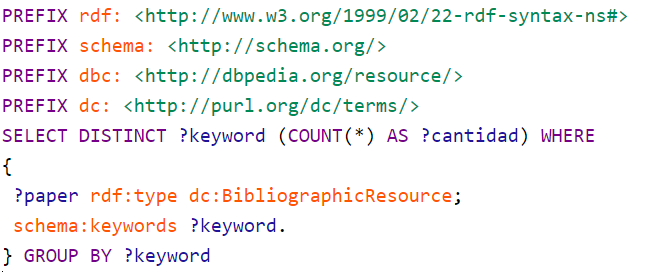


**Resultado**:

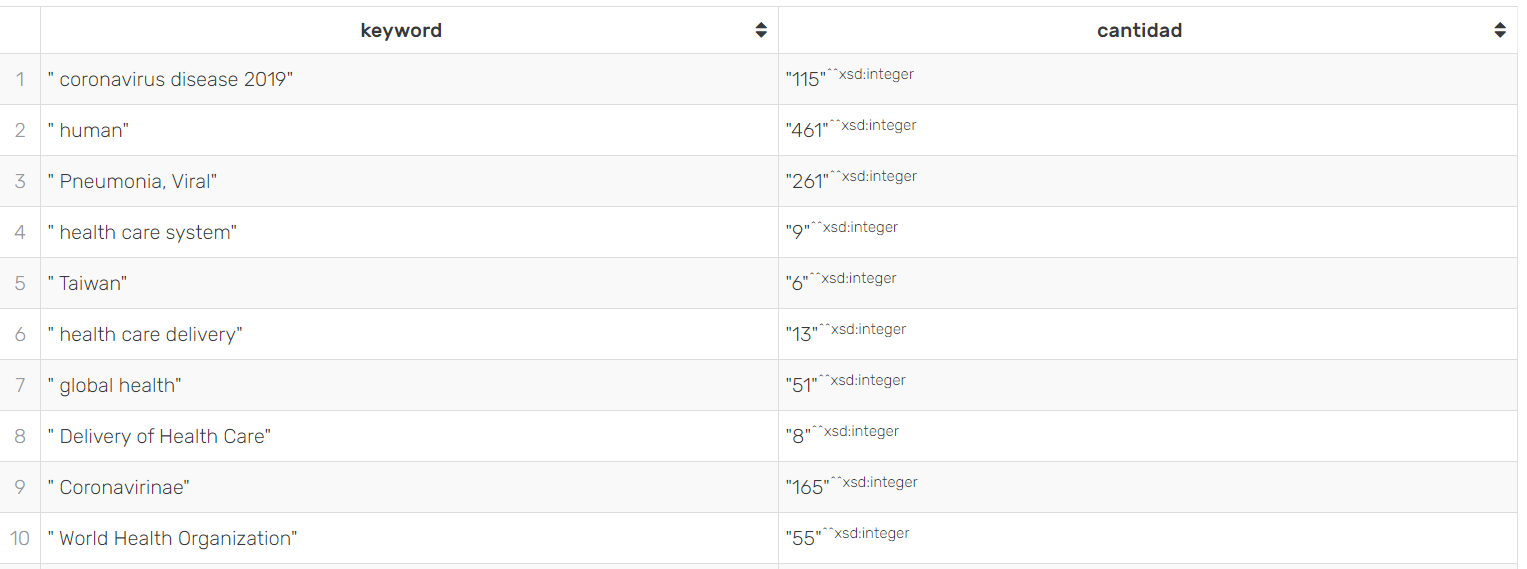


**Pregunta:** consulta sobre la cantidad de las palabras claves que aparecen en las publicaciones de covid-19

Consulta:

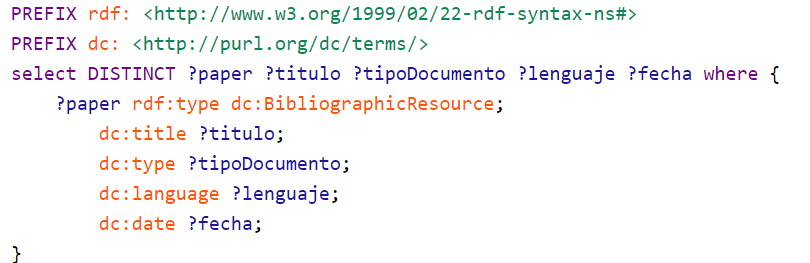


**Resultado:**

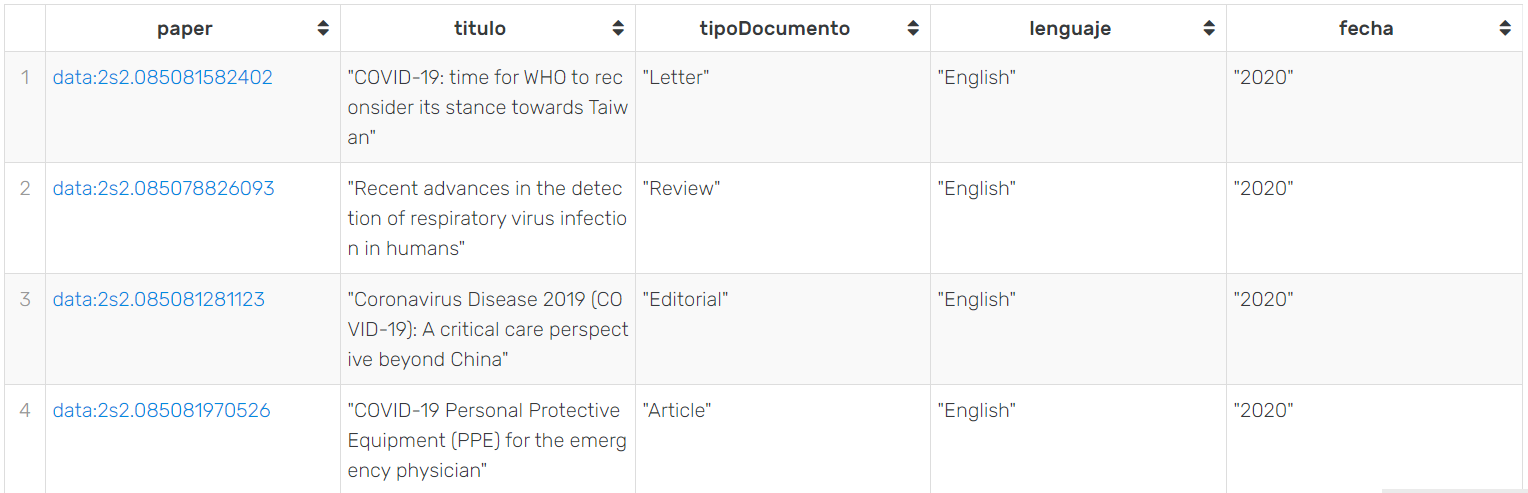


**Pregunta:** consultar las propiedades principales de todos los papers.

**Consulta**:

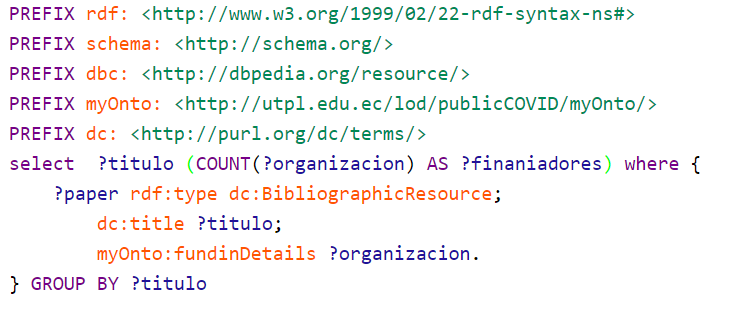


**Resultado**:

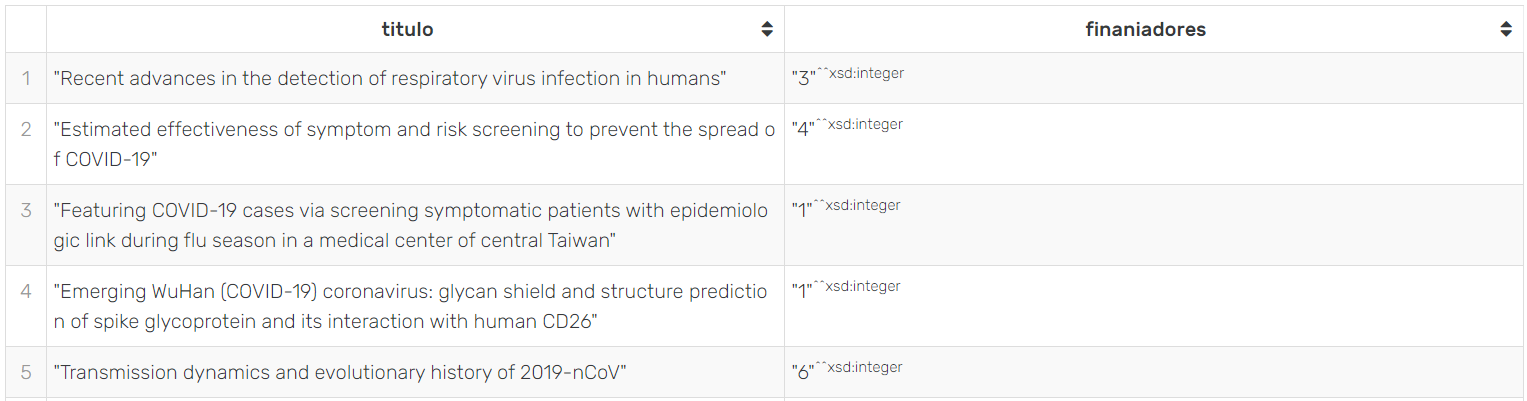


**Pregunta:** cual es la cantidad de financiadores por publicaciones.

**Consulta**:

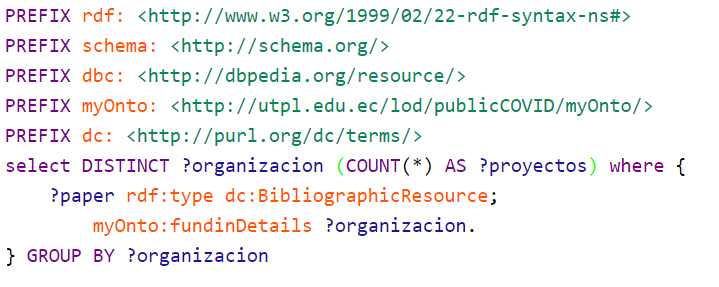


**Resultado**:

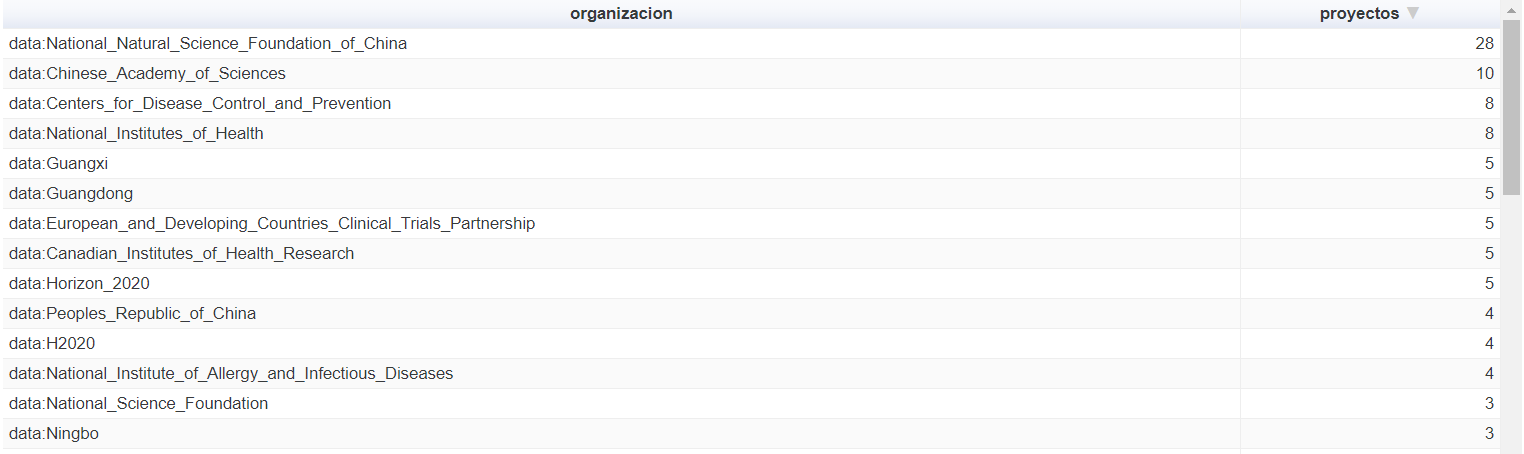


**Pregunta:** cual es la organización que más a finalizado las investigaciones de Covid-19.

**Consulta**:



**Resultado**:



# Aplicación propuesta

En esta sección se describe las tecnologías y funcionamiento general de la aplicación propuesta. Para obtener los datos de la base GrahpDB se usó una librería de Java con el nombre de “rdf4j”. Y para podr usar dichos datos dse debe generar la aplicación web mediante, y esto se lo hizo por medio del framework sparkjava. Los resultados son los siguientes.



# Conclusiones

* La web semántica nos permite crear una posibilidad abierta de estructurar los datos que encontramos en la web con la finalidad de darle una estructura y sentido a la data para transformarla a conocimiento e información que puede ser usada por muchas personas comunes, profesionales o investigadores para dar soluciones al mundo real.
* Jena es un framework que nos facilita la creación de datos RDF de manera sencilla y rápida.
* GraphDB es un motor de base de datos semántica orientada a grafos lo que permite el fácil entendimiento de los datos a almacenar.
* Con los avances tecnológicos se puede llevar acabo un sinnúmero de aplicaciones de generadores de conocimiento, esto abre muchas más posibilidades para crear una web semántica de formas mucho más sencillas y rápidas de acoplarlas. Esto incentiva abrir mas posibilidades para mostrar el conocimiento a cualquier persona, sociedades u organizaciones interesadas en el campo del Big Data o Análisis de datos.

# Bibliografía

Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in big data analytics. *Journal of parallel and distributed computing*, *74*(7), 2561-2573.

Zhu, Y., Wan, J., Zhou, Z., Chen, L., Qiu, L., Zhang, W., ... & Yu, Y. (2019, July). Triple-to-Text: Converting RDF Triples into High-Quality Natural Languages via Optimizing an Inverse KL Divergence. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 455-464).