



UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA

Sistemas Basados en Conocimientos

Tema: Análisis y recolección de datos

Estudiante: Alexis Fabian Montoya Parra

Juan Andrés Ramon Zhigui

Luis Miguel Rojas Paccha

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 Crossref y Scholar Academic, una base de datos científica en donde se puede encontrar 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 más 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.

	B	C	D	E	F	G
1	date-time	publisher	DOI	type	source	title
3	2020-10-23T17:13:59Z	The International Centre for Trade Unio	10.14213/inteuniorigh.27.3.0020	journal-article	Crossref	'Just because you don't see your boss, doesn't mean you don't have a boss': Covic
4	2020-08-14T01:40:41Z	American Association for the Advancem	10.1126/science.abe2995	journal-article	Crossref	'We're losing an entire generation of scientists.' COVID-19's economic toll hits La
10	2020-11-17T07:44:34Z	Elsevier BV	10.1016/j.outlook.2020.08.013	journal-article	Crossref	Aging in America: How COVID-19 Will Change Care, Coverage, and Compassion
13	2020-07-02T14:40:29Z	Middle Atlantic Review of Latin America	10.23870/marlas.312	journal-article	Crossref	America Latina y el Covid-19
14	2020-12-21T22:17:22Z	Editorial Universidad de Sevilla	10.12795/araucaria.2020.i45.16	journal-article	Crossref	América Latina y la Unión Europea: agendas sociales, competencia geopolítica y t
15	2020-12-08T05:21:13Z	Universidad Nacional Autonoma de Me	10.22201/fe.18701442e.2020.37.77770	journal-article	Crossref	América Latina: Hacia un periodo de débil crecimiento y COVID-19
16	2021-01-15T21:23:47Z	Elsevier BV	10.1016/j.explore.2020.08.012	journal-article	Crossref	America, consciousness, COVID-19, climate change, and migration
20	2021-03-29T14:24:45Z	Oxford University Press (OUP)	10.1093/jtm/taaa176	journal-article	Crossref	Analysis of travel restrictions for COVID-19 control in Latin America through netw
22	2020-07-26T13:40:28Z	Project Muse	10.1353/lag.2020.0049	journal-article	Crossref	As organizações religiosas brasileiras frente à pandemia de Covid-19
23	2021-01-25T05:05:46Z	Project Muse	10.1353/lag.2020.0057	journal-article	Crossref	Asylum and Mass Detention at the U.S.-Mexico Border during Covid-19
24	2020-09-10T02:11:21Z	Human Kinetics	10.1123/ijsc.2020-0217	journal-article	Crossref	Australian Football in America During COVID-19
26	2021-05-09T11:41:00Z	Ovid Technologies (Wolters Kluwer Heal	10.1097/ede.0000000000001293	journal-article	Crossref	Being a Latin American Woman in Science During the COVID-19 Pandemic
27	2020-11-16T14:47:30Z	Clinical Biotec	10.21931/rb/2020.05.04.27	journal-article	Crossref	Bioethical Guidelines of 'Extreme Triage' Under Covid: The Question of 'Possible l
30	2020-10-29T16:24:57Z	BMJ	10.1136/bmjhci-2020-100159	journal-article	Crossref	Collaboration in times of COVID-19: the urgent need for open-data sharing in Lati
34	2020-09-02T12:26:39Z	SAGE Publications	10.1177/0003134820927313	journal-article	Crossref	Coronavirus Disease 2019 (COVID-19) and Surgical Recommendations in Latin Am
35	2020-10-16T08:43:43Z	Gavin Publishers	10.29011/2690-9480.100122	journal-article	Crossref	Corporate Responses to COVID-19 Layoffs in North America and the Role of Hum
36	2020-08-03T07:17:11Z	International Institute for Science, Tech	10.7176/jlpg/99-13	journal-article	Crossref	COVID -19 Emergency Laws and Law Enforcement in Nigeria America And Britain.
37	2021-02-22T15:43:20Z	Informa UK Limited	10.1080/07399332.2020.1833884	journal-article	Crossref	COVID 19: sexual vulnerabilities and gender perspectives in Latin America
38	2021-02-08T21:16:49Z	Wiley	10.1111/blar.13180	journal-article	Crossref	COVID -19 and Historical Global Rupture in Latin America
39	2021-02-08T20:54:36Z	Wiley	10.1111/blar.13188	journal-article	Crossref	COVID -19 and the Limitations of Official Responses to Gender-Based Violence in

Selección de URIS

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

Prefijo	URI	Tipo
dataPrefix	http://ky.utpl.edu.ec/publicicovid/data	Definida
EventPrefix	http://purl.org/NET/c4dm/event.owl	Establecida en la web
CPrefix	http://purl.org/spar/c4o/	Establecida en la web
vcard	http://www.w3.org/2006/vcard/ns	Establecida en la web
foaf	http://xmlns.com/foaf/0.1/	Establecida en la web
dbo	http://dbpedia.org/ontology/	Establecida en la web
vivo	http://vivoweb.org/ontology/core	Establecida en la web
bibo	http://purl.org/ontology/bibo	Establecida en la web
dct	http://purl.org/dc/terms/	Establecida en la web
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns	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 realizó 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 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 respuestas mediante un REQUEST (Ver la siguiente figura).

```

Archivo Edición Formato Ver Ayuda
["101055a-1229-5048", "http://dbpedia.org/resource/ISWC", 0.24496769905090332, 0.31707316637039185, "http://dbpedia.org/resource/Category:Music_technology"]
["101055a-1229-5048", "http://dbpedia.org/resource/ISWC", 0.24496769905090332, 0.31707316637039185, "http://dbpedia.org/resource/Category:ISO_standards"]
["101055a-1229-5048", "http://dbpedia.org/resource/ISWC", 0.24496769905090332, 0.31707316637039185, "http://dbpedia.org/resource/Category:Identifiers"]
["101055a-1229-5048", "http://dbpedia.org/resource/ISWC", 0.24496769905090332, 0.31707316637039185, "http://dbpedia.org/resource/Category:Universal_identifiers"]
["101055a-1229-5048", "http://dbpedia.org/resource/RDF", 0.3218223750591278, 0.4697193503379822, "http://dbpedia.org/resource/Category:Resource_Description_Framework"]
["101055a-1229-5048", "http://dbpedia.org/resource/RDF", 0.3218223750591278, 0.4697193503379822, "http://dbpedia.org/resource/Category:Knowledge_representation"]
["101055a-1229-5048", "http://dbpedia.org/resource/RDF", 0.3218223750591278, 0.4697193503379822, "http://dbpedia.org/resource/Category:World_Wide_Web_Consortium_standards"]
["101055a-1229-5048", "http://dbpedia.org/resource/RDF", 0.3218223750591278, 0.4697193503379822, "http://dbpedia.org/resource/Category:XML"]
["101055a-1229-5048", "http://dbpedia.org/resource/RDF", 0.3218223750591278, 0.4697193503379822, "http://dbpedia.org/resource/Category:XML-based_standards"]
["101055a-1229-5048", "http://dbpedia.org/resource/RDF", 0.3218223750591278, 0.4697193503379822, "http://dbpedia.org/resource/Category:Metadata"]
["101055a-1229-5048", "http://dbpedia.org/resource/RDF", 0.3218223750591278, 0.4697193503379822, "http://dbpedia.org/resource/Category:Semantic_Web"]
["101055a-1229-5048", "http://dbpedia.org/resource/RDF", 0.3218223750591278, 0.4697193503379822, "http://dbpedia.org/resource/Category:Bibliography_file_formats"]
["101055a-1229-5048", "http://dbpedia.org/resource/RDF_triples", 0.19807380437850952, 0.222222238779068, "http://dbpedia.org/resource/Category:Resource_Description_Framework"]
["101055a-1229-5048", "http://dbpedia.org/resource/RDF_triples", 0.19807380437850952, 0.222222238779068, "http://dbpedia.org/resource/Category:Knowledge_representation"]
["101055a-1229-5048", "http://dbpedia.org/resource/RDF_triples", 0.19807380437850952, 0.222222238779068, "http://dbpedia.org/resource/Category:World_Wide_Web_Consortium_standards"]
["101055a-1229-5048", "http://dbpedia.org/resource/RDF_triples", 0.19807380437850952, 0.222222238779068, "http://dbpedia.org/resource/Category:XML"]
["101055a-1229-5048", "http://dbpedia.org/resource/RDF_triples", 0.19807380437850952, 0.222222238779068, "http://dbpedia.org/resource/Category:XML-based_standards"]
["101055a-1229-5048", "http://dbpedia.org/resource/RDF_triples", 0.19807380437850952, 0.222222238779068, "http://dbpedia.org/resource/Category:Metadata"]
["101055a-1229-5048", "http://dbpedia.org/resource/RDF_triples", 0.19807380437850952, 0.222222238779068, "http://dbpedia.org/resource/Category:Semantic_Web"]
["101055a-1229-5048", "http://dbpedia.org/resource/RDF_triples", 0.19807380437850952, 0.222222238779068, "http://dbpedia.org/resource/Category:Bibliography_file_formats"]
["101055a-1229-5048", "http://dbpedia.org/resource/triples", 0.17579340934753418, 0.35158681869506836, "http://dbpedia.org/resource/Category:Baseball_terminology"]
["101055a-1229-5048", "http://dbpedia.org/resource/triples", 0.17579340934753418, 0.35158681869506836, "http://dbpedia.org/resource/Category:Baseball_statistics"]
["101055a-1229-5048", "http://dbpedia.org/resource/knowledge_graph", 0.632880449295044, 0.9714285731315613, "http://dbpedia.org/resource/Category:Knowledge_engineering"]
["101055a-1229-5048", "http://dbpedia.org/resource/knowledge_graph", 0.632880449295044, 0.9714285731315613, "http://dbpedia.org/resource/Category:Technical_communication"]
["101055a-1229-5048", "http://dbpedia.org/resource/knowledge_graph", 0.632880449295044, 0.9714285731315613, "http://dbpedia.org/resource/Category:Information_science"]
["101055a-1229-5048", "http://dbpedia.org/resource/knowledge_graph", 0.632880449295044, 0.9714285731315613, "http://dbpedia.org/resource/Category:Semantic_Web"]
["101055a-1229-5048", "http://dbpedia.org/resource/knowledge_graph", 0.632880449295044, 0.9714285731315613, "http://dbpedia.org/resource/Category:Ontology_(information_science)"]
["101055a-1229-5048", "http://dbpedia.org/resource/knowledge_graph", 0.632880449295044, 0.9714285731315613, "http://dbpedia.org/resource/Category:Knowledge_representation"]
["101055a-1229-5048", "http://dbpedia.org/resource/DrugBank", 0.5851426887512207, 1, "http://dbpedia.org/resource/Category:Chemical_databases"]
["101055a-1229-5048", "http://dbpedia.org/resource/DrugBank", 0.5851426887512207, 1, "http://dbpedia.org/resource/Category:Metabolic_databases"]
["101055a-1229-5048", "http://dbpedia.org/resource/GERBIL", 0.31589147448539734, 0.6317829489707947, "http://dbpedia.org/resource/Category:Muridae"]
["101055a-1229-5048", "http://dbpedia.org/resource/GERBIL", 0.31589147448539734, 0.6317829489707947, "http://dbpedia.org/resource/Category:Gerbils"]
["101055a-1229-5048", "http://dbpedia.org/resource/GERBIL", 0.31589147448539734, 0.6317829489707947, "http://dbpedia.org/resource/Category:Pet_rodents"]
["101055a-1229-5048", "http://dbpedia.org/resource/GERBIL", 0.31589147448539734, 0.6317829489707947, "http://dbpedia.org/resource/Category:Fauna_of_Central_Asia"]
["101055a-1229-5048", "http://dbpedia.org/resource/GERBIL", 0.31589147448539734, 0.6317829489707947, "http://dbpedia.org/resource/Category:Fauna_of_Mongolia"]
["101055a-1229-5048", "http://dbpedia.org/resource/ISWC", 0.2306729555130005, 0.31707316637039185, "http://dbpedia.org/resource/Category:Music_technology"]
["101055a-1229-5048", "http://dbpedia.org/resource/ISWC", 0.2306729555130005, 0.31707316637039185, "http://dbpedia.org/resource/Category:ISO_standards"]
["101055a-1229-5048", "http://dbpedia.org/resource/ISWC", 0.2306729555130005, 0.31707316637039185, "http://dbpedia.org/resource/Category:Identifiers"]
["101055a-1229-5048", "http://dbpedia.org/resource/ISWC", 0.2306729555130005, 0.31707316637039185, "http://dbpedia.org/resource/Category:Universal_identifiers"]

```

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

Una vez obtenido todos los datos necesarios se procedió 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.

```

tor Run Debug Profile Team Tools Window Help
Search (Ctrl+F)

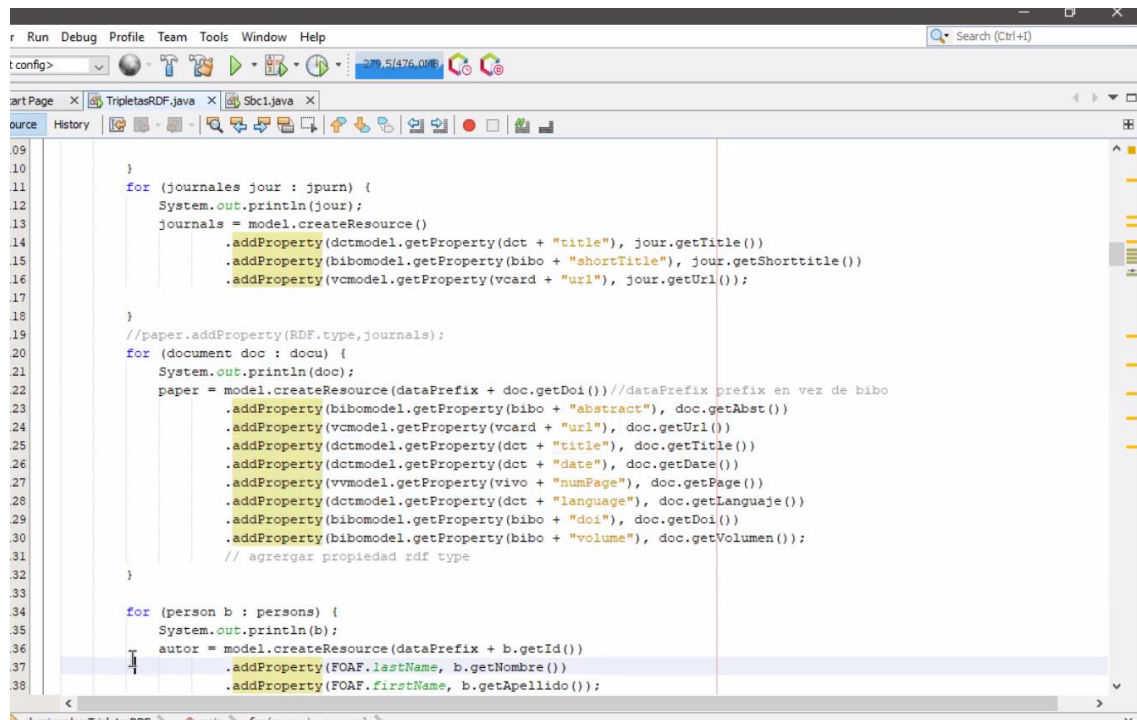
ult config>
231.9/47 DMB

Start Page X TripletasRDF.java X Sbci.java X
Source History
73 model.setNsPrefix("c4o", CPrefix);
74 Model cpmodel = ModelFactory.createDefaultModel();
75
76 String vcard = "http://www.w3.org/2006/vcard/ns#";
77 model.setNsPrefix("vcard", vcard);
78 Model vcmodel = ModelFactory.createDefaultModel();
79
80 String foaf = "http://xmlns.com/foaf/0.1/";
81 model.setNsPrefix("foaf", foaf);
82 Model foafmodel = ModelFactory.createDefaultModel();
83
84 String dbo = "http://dbpedia.org/ontology/";
85 model.setNsPrefix("dbo", dbo);
86 Model dboModel = ModelFactory.createDefaultModel();
87
88 String vivo = "http://vivoweb.org/ontology/core#";
89 model.setNsPrefix("vivo", vivo);
90 Model vmmodel = ModelFactory.createDefaultModel();
91
92 String bibo = "http://purl.org/ontology/bibo#";
93 model.setNsPrefix("bibo", bibo);
94 Model bibomodel = ModelFactory.createDefaultModel();
95
96 String dct = "http://purl.org/dc/terms/";
97 model.setNsPrefix("dct", dct);
98 Model dctmodel = ModelFactory.createDefaultModel();
99
100 String rdf = "http://www.w3.org/1999/02/22-rdf-syntax-ns#";
101 model.setNsPrefix("rdf", rdf);
102 Model rdfmodel = ModelFactory.createDefaultModel();

```

Después se realizó 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.

A screenshot of an IDE window showing a Java file named 'Sbc1.java'. The code is generating RDF triples for a set of journals and a document. It uses several models (bibomodel, dctmodel, vcard, vvmmodel) to add properties to resources. The code is structured with 'for' loops to iterate over collections of journals, documents, and persons. The IDE interface includes a menu bar (File, Run, Debug, Profile, Team, Tools, Window, Help), a toolbar, and a search bar. The code is as follows:

```
09  
10  
11 for (journals jour : jpurn) {  
12     System.out.println(jour);  
13     journals = model.createResource()  
14         .addProperty(dctmodel.getProperty(dct + "title"), jour.getTitle())  
15         .addProperty(bibomodel.getProperty(bibo + "shortTitle"), jour.getShorttitle())  
16         .addProperty(vcardmodel.getProperty(vcard + "url"), jour.getUrl());  
17  
18 }  
19 //paper.addProperty(RDF.type,journals);  
20 for (document doc : docu) {  
21     System.out.println(doc);  
22     paper = model.createResource(dataPrefix + doc.getDoi())//dataPrefix prefix en vez de bibo  
23         .addProperty(bibomodel.getProperty(bibo + "abstract"), doc.getAbst())  
24         .addProperty(vcardmodel.getProperty(vcard + "url"), doc.getUrl())  
25         .addProperty(dctmodel.getProperty(dct + "title"), doc.getTitle())  
26         .addProperty(dctmodel.getProperty(dct + "date"), doc.getDate())  
27         .addProperty(vvmmodel.getProperty(vivo + "numPage"), doc.getPage())  
28         .addProperty(dctmodel.getProperty(dct + "language"), doc.getLanguaje())  
29         .addProperty(bibomodel.getProperty(bibo + "doi"), doc.getDoi())  
30         .addProperty(bibomodel.getProperty(bibo + "volume"), doc.getVolumen());  
31     // agregar propiedad rdf type  
32 }  
33  
34 for (person b : persons) {  
35     System.out.println(b);  
36     autor = model.createResource(dataPrefix + b.getId())  
37         .addProperty(FOAF.lastName, b.getNombre())  
38         .addProperty(FOAF.firstName, b.getApellido());
```

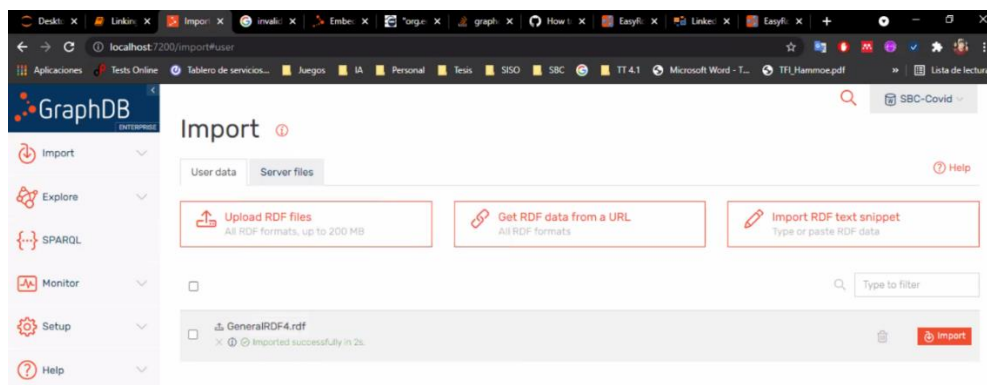
En la siguiente figura se puede apreciar un extracto sobre las tripletas generadas por Jena.

```
GeneralRDF4: Bloc de notas
Archivo Edición Formato Ver Ayuda

<rdf:RDF
  xmlns:dct="http://purl.org/dc/terms/"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:c4o="http://purl.org/spar/c4o/"
  xmlns:event="http://purl.org/NET/c4dm/event.owl#"
  xmlns:dbo="http://dbpedia.org/ontology/"
  xmlns:vcard="http://www.w3.org/2006/vcard/ns#"
  xmlns:data="http://ky.utpl.edu.ec/publicicovid/data#"
  xmlns:bibo="http://purl.org/ontology/bibo#"
  xmlns:vivo="http://vivoweb.org/ontology/core#"
  xmlns:foaf="http://xmlns.com/foaf/0.1/"
  <rdf:Description>
    <vcard:url>N/A</vcard:url>
    <bibo:shortTitle>N/A</bibo:shortTitle>
    <dct:title>&lt;strong&gt;&lt;/strong&gt;COVID-19: Prediction of Vulnerable Areas Having High
  </rdf:Description>
  <rdf:Description>
    <vcard:url>N/A</vcard:url>
    <bibo:shortTitle>N/A</bibo:shortTitle>
    <dct:title>Drivers and Barriers to Living in a Multigenerational Household Pre-COVID - Mid-
  </rdf:Description>
  <rdf:Description>
    <vcard:url>N/A</vcard:url>
    <bibo:shortTitle>Social Work and the COVID-19 Pandemic</bibo:shortTitle>
    <dct:title>COVID-19 and Social Inequalities: A Political View From Social Work</dct:title>
  </rdf:Description>
  <rdf:Description rdf:about="http://ky.utpl.edu.ec/publicicovid/data#191">
    <foaf:lastName>Thiago</foaf:lastName>
    <vcard:url>http://dx.doi.org/10.4324/9781003154037-19</vcard:url>
    <bibo:doi>191</bibo:doi>
    <data:doiDocu>1043249781003154037-21</data:doiDocu>
    <data:idpersona>191</data:idpersona>
    <bibo:volume>N/A</bibo:volume>
    <dct:title>The US</dct:title>
    <vivo:numPage> and Covid-19</vivo:numPage>
    <dct:language>5/4/2021</dct:language>
```

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 más gestionado sobre la información de las publicaciones, por ello se usó la base de grafos denominada “GraphDB”. En la siguiente imagen se puede apreciar los datos subidos a la base de datos.



Consultas SPARQL

En si 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: consultar las propiedades principales de todos los papers.

Consulta:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dc: <http://purl.org/dc/terms/>
select DISTINCT ?paper ?titulo ?tipoDocumento ?lenguaje ?fecha where {
    ?paper rdf:type dc:BibliographicResource;
        dc:title ?titulo;
        dc:type ?tipoDocumento;
        dc:language ?lenguaje;
        dc:date ?fecha;
}
```

Resultado:

	paper	titulo	tipoDocumento	lenguaje	fecha
1	data:2s2.085081582402	"COVID-19: time for WHO to reconsider its stance towards Taiwan"	"Letter"	"English"	"2020"
2	data:2s2.085078826093	"Recent advances in the detection of respiratory virus infection in humans"	"Review"	"English"	"2020"
3	data:2s2.085081281123	"Coronavirus Disease 2019 (COVID-19): A critical care perspective beyond China"	"Editorial"	"English"	"2020"
4	data:2s2.085081970526	"COVID-19 Personal Protective Equipment (PPE) for the emergency physician"	"Article"	"English"	"2020"

Aplicación:

Para compilar la aplicación tomar en cuenta lo siguiente:

1. Instalar **node, yarn, npm**
2. Descargar las dependencias: **yarn install**
3. Compilar: **yarn start**

Par la visualización de las consultas únicamente se hace clic en los nombres en rojo, cada uno de los ítems descritos es una consulta Sparql.

Se muestra a continuación:



Para el consumo del endpoint se utiliza el siguiente que está en línea:
<http://graphdb.linked-open-statistics.org/repositories>

```
const MAIN_ENDPOINT = 'http://graphdb.linked-open-statistics.org/repositories';

export default {
  TEST_ENDPOINT: `${MAIN_ENDPOINT}/test`,
  COVID_ENDPOINT: `${MAIN_ENDPOINT}/CovidLatam`,
};

export const TOKEN = process.env.REACT_APP_MGL_TOKEN;
```

Enlace Github: <https://github.com/Jtaramon/ProyectoSbc>