Sistemas Basados en Conocimientos

Integrante:

- Bryant Portilla
- Diego Pinto

Tema:

- Bibliometric Data

Fecha:

- 26/06/2020

Docente:

- Ing. Janeth Chicaiza

Periodo

Abril/2020 - Agosto/2020

Tabla resumen de datos recolectados

Class	Links
myData:Author	14K
myData:Article	9K
fabio:Review	3K
fabio:Letter	791
myData:Editorial	463
myData:Note	412
fabio:ConferencePaper	100
myData:ShortSurvey	83
fabio:Erratum	20
myData:DataPaper	16

Pre-procesamiento de datos

Teniendo todas estas características de estos datos, se puede pensar en las principales actividades requeridas para la data que se basa en la extracción, transformación y visualización, este método se lo denomina ETL (Extra, transforma y carga).

La extracción se encarga de la recopilación de datos desde la plataforma de la base de datos científica SCOPUS en la cuales se encuentran en un formato CSV. En la transformación se encarga de convertir estos datos a un formato (*tripleta*) más fácil de mantener y trabajar con grandes conjuntos de datos. Y en la visualización o carga se busca presentar los datos en sitio web semántico.

En la extracción se usó estas cadenas de búsquedas en la base de datos científica para obtener los resultados que se va a transformar:

- **TITLE-ABS-KEY ((covid-19) AND (sars-cov-2)) AND (LIMIT-TO (PUBSTAGE , "final"))**
-

- **TITLE-ABS-KEY ((covid-19) AND (sars-cov-2)) AND (LIMIT-TO (PUBSTAGE , "aip"))**

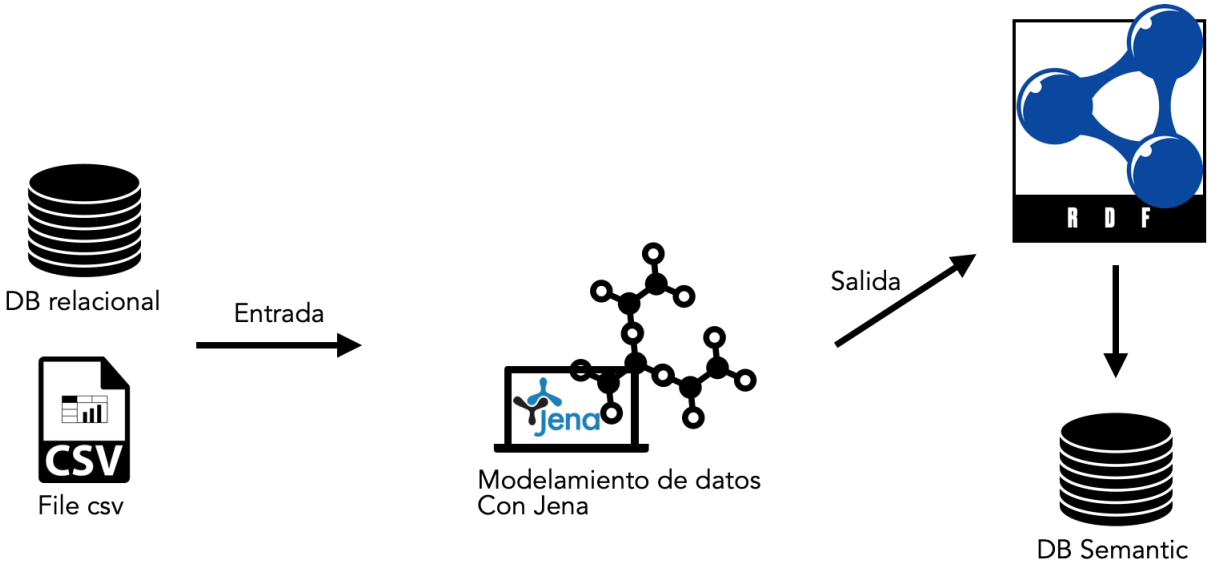
Siguiendo el método ETL, una vez extraídos se procede la parte más fuerte que es la transformación de los datos, en esta parte se realiza múltiples actividades en nuestro caso se hizo la limpieza y validación de los datos que se extrajo y también buscamos duplicidad de información que existan en el dataset principal.

En la Figura 1 se observa un fragmento de los datos ya transformados.

idArticulo	Authors	Authors_ID	Title	EID	Year	Sourc	Volum	Issue	Art. Ni	Pa	Page end	Cited by	DOI
BiblioData	Byass P.	7006527895;	Eco-epidemiological asses	2-s2.0-8508	2020	Global H	13	1	1760490				10.1080/165
BiblioData	Liu P., Cai J., Jia R., Xia S.	57200337904;572	Dynamic surveillance of SA	2-s2.0-8508	2020	Emergin	9	1		1254	1258		10.1080/222
BiblioData	Islam H., Rahman A., Ma	57195434373;572	A generalized overview of	2-s2.0-8508	2020	Electroni	17	6	em251				10.29333/ejg
BiblioData	Wang C., Li W., Drabek D	57216967065;572	A human monoclonal antil	2-s2.0-8508	2020	Nature C	11	1	2251			12	10.1038/s41
BiblioData	Lau S.-Y., Wang P., Mok I	57208238618;549	Attenuated SARS-CoV-2 va	2-s2.0-8508	2020	Emergin	9	1		837	842	1	10.1080/222
BiblioData	Beloncle F.M., Pavlovsky	57216557219;572	Recruitability and effect o	2-s2.0-8508	2020	Annals o	10	1	55				10.1186/s13
BiblioData	Grimaud M., Starck J., Le	55043167800;572	Acute myocarditis and mu	2-s2.0-8508	2020	Annals o	10	1	69				10.1186/s13
BiblioData	Eslami H., Jalili M.	6506092327;5721	The role of environmental	2-s2.0-8508	2020	AMB Exp	10	1	92				10.1186/s13
BiblioData	Suo T., Liu X., Feng J., Gu	57193390916;572	ddPCR: a more accurate to	2-s2.0-8508	2020	Emergin	9	1		1259	1268		10.1080/222
BiblioData	Yang X., Dong N., Chan E	57217073220;572	Genetic cluster analysis of	2-s2.0-8508	2020	Emergin	9	1		1287	1299		10.1080/222
BiblioData	Maitra A., Sarkar M.C., R	57217125767;572	Mutations in SARS-CoV-2	2-s2.0-8508	2020	Journal o	45	1	76				10.1007/s12

Figura 1. *Datos limpios.*

Transformación de datos:



Git de la aplicación:

https://github.com/diepinto30/GenerateBibliometricDataCovid

Resultada de una búsqueda con los datos en GraphDB