Universidad Técnica Particular de Loja

Informe Final

Componente:
Sistemas Basados en el Conocimiento

Integrantes:
Jean Paul Mosquera
Marco Caicedo

Fecha:
14 de julio del 2020

Loja – Ecuador

Proceso
(Obtención, Limpieza, Enlazado de datos)

1. Definición de URIs

Para definir las URIs previamente se revisó la documentación compartida en función a las indicaciones se ha definido la uri base que es la siguiente

A. 2.3.1 Especificación

- **Identificación De fuentes y Licencias**

Id	Fuente	Licencia
1	https://www.datos.gov.co/Salud-y-Protecci-n-Social/Casos-positivos-de-COVID-19-en-Colombia/gt2j-8ykr/data	Open
2	https://github.com/andrab/ecuacovid	Open
3	https://www.acaps.org/covid19-government-measures-dataset	Creative Commons
4	https://data.humdata.org/dataset/total-covid-19-tests-performed-by-country	Creative Commons
5	https://github.com/owid/covid-19-data/tree/master/public/data/	Creative Commons
14	https://github.com/microsoft/Bing-COVID-19-Data/tree/master/data	Creative Commons

- **Diseño de la URI**

Primeramente, se ha considerado definir las uri base que describirán tanto el modelo ontológico, como las instancias las cuales se describen a continuación.

Ontología (Clases/Relaciones) https://ld.utpl.edu.ec/dataCOVID/ontology#

Datos (Instancias/Individuos) https://ld.utpl.edu.ec/dataCOVID/resource/

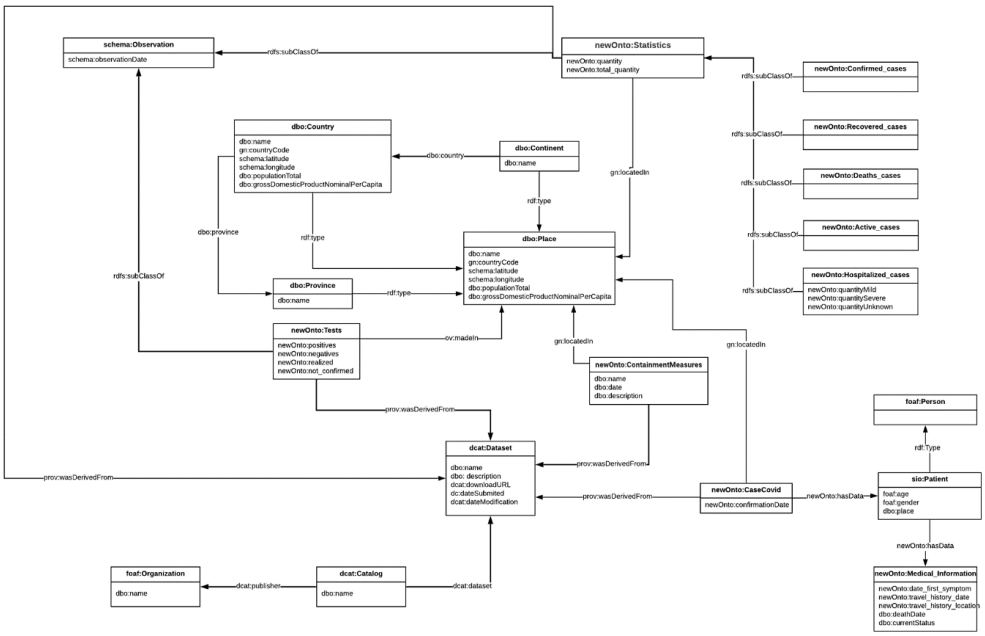
Posteriormente se ha considerado emplear Patterned URIs, puesto esto permitirá a los usuarios encontrar fácilmente y de forma organizadas los recursos, para agrupar los datos en función a su clase así se tendrían las siguientes URIs:

Clase| Categoria	Uri
Place	https://ld.utpl.edu.ec/dataCOVID/data/[Place] **Uri base Continente**

	https://ld.utpl.edu.ec/dataCOVID/data/continent/[ISO-3166] Ejemplo Continente Sur América https://ld.utpl.edu.ec/dataCOVID/data/continent/SA **Uri base Pais** https://ld.utpl.edu.ec/dataCOVID/data/country/[ISO-3166-Nivel-1] Ejemplo País Ecuador https://ld.utpl.edu.ec/dataCOVID/data/country/EC **Uri base Provincia** https://ld.utpl.edu.ec/dataCOVID/data/province/[ISO-3166-Nivel-2] Ejemplo Provincia Loja https://ld.utpl.edu.ec/dataCOVID/data/province/EC-LO
Statistic	https://ld.utpl.edu.ec/dataCOVID/data/statistic/[Statistic] **Ejemplo casos positivos** https://ld.utpl.edu.ec/dataCOVID/data/statistic/positive-cases
Tests	https://ld.utpl.edu.ec/dataCOVID/data/tests/[Tests] **Ejemplo casos positivos** https://ld.utpl.edu.ec/dataCOVID/data/tests/realized
Dataset	https://ld.utpl.edu.ec/dataCOVID/data/dataset/[Proveedor_Dataset] Ejemplo datatset

	https://ld.utpl.edu.ec/dataCOVID/data/dataset/Bing
Pacient	https://ld.utpl.edu.ec/dataCOVID/data/pacient/[ISO-3166-Nivel-2]/[Id_Pacient] Ejemplo Paciente de Antoquia Colombia https://ld.utpl.edu.ec/dataCOVID/data/pacient/CO-ANT/1
Medical_Information	https://ld.utpl.edu.ec/dataCOVID/data/medical_information/[ISO-3166-Nivel-2]/[Id_ Medical_Information] Ejemplo información médica del Paciente 1 de Antoquia Colombia https://ld.utpl.edu.ec/dataCOVID/data/medical_information /CO-ANT/1
Goverment_Action	https://ld.utpl.edu.ec/dataCOVID/data/goverment_action/[ISO-3166-Nivel-1]/[Id_Goverment_Action] Ejemplo acción gubernamental de Ecuador https://ld.utpl.edu.ec/dataCOVID/data/ goverment_action/

B. 2.3.2 Modelado



C. 2.3.3 Generación

Para la parte de generación, que comprende la transformación, limpieza y enlazado, se está considerando la aplicación de la siguiente estrategia, que consiste en:

- **Extracción de todos los datos de los diferentes dataset antes analizados.**
- **Limpieza y disposición de datos en otro fichero**
- **Filtrado de datos en función a país con el objetivo de obtener únicamente datos de los países latinoamericanos.**
- **Filtrado de campos acorde a los requeridos en las clases del modelo ontológico**
- **Creación de Script para lectura de fichero y posterior uso de los datos.**

2. **Actualización de datos**

Antes de realizar la actualización de datos se ha procedido a validar la última actualización de datos a la que fue sometida la información contenida en la fuente de datos.

Id	Fuente	Tipo	Data	Ultima Actualizaci ón Fuente
1	https://www.datos.gov.co/Salud-y-Protecci-n-Social/Casos-positivos-de-COVID-19-en-Colombia/gt2j-8ykr/data	CSV	Colombia	20/06/20
2	https://github.com/andrab/ecuacovid	CSV	Ecuador	21/06/20
3	https://data.humdata.org/dataset/total-covid-19-tests-performed-by-country	CSV	Latam	14/06/20
4	https://www.acaps.org/covid19-government-measures-dataset	CSV	Government Actios	18/06/20
5	https://github.com/owid/covid-19-data/tree/master/public/data/	CSV	Latam/World	21/06/20
6	https://github.com/microsoft/Bing-COVID-19-Data/tree/master/data	CSV	Latam/Wordl	21/06/20

Con ello se ha procedido a realizar la actualización de las fuentes de datos con la información más reciente.

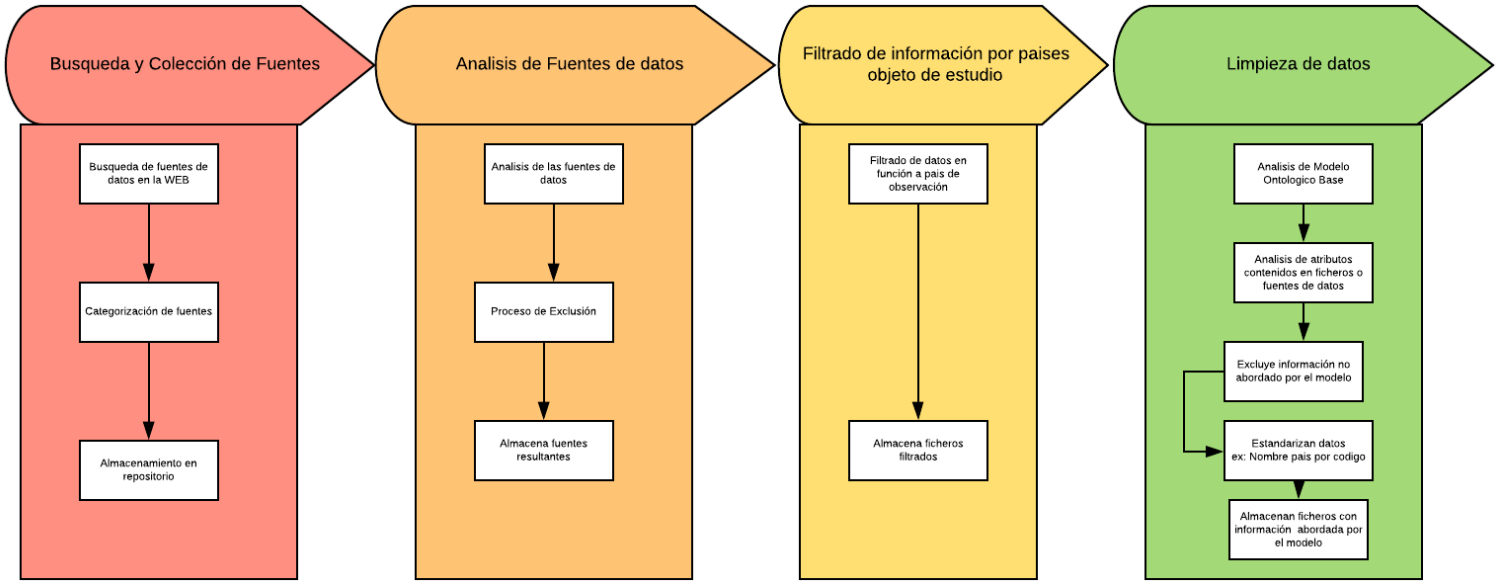
Se ha procedido a analizar métodos para adaptar los datos al modelo ontológico propuesto previo a la generación de individuos o instancias empleando Jena.

De este modo se ha considerado en función a los atributos previamente definidos clasificar las fuentes de datos en función a la data que proveen, sean estas estadísticas, medidas o acciones gubernamentales, descripción de casos, entre otras, para de este modo lograr un proceso más limpio al momento de adaptar las fuentes de datos al modelo propuesto. Así se ha definido la información a obtener de cada una de las fuentes para evitar redundancia.

Id	Fuente	Data
1	https://www.datos.gov.co/Salud-y-Protecci-n-Social/Casos-positivos-de-	INFORMACION MEDICA DE CASOS EN COLOMBIA

	COVID-19-en-Colombia/gt2j-8ykr/data	
2	https://github.com/andrab/ecuacovid	ESTADISTICAS HOSPITALIZADOS, MUESTRAS
3	https://data.humdata.org/dataset/total-covid-19-tests-performed-by-country	TESTS
4	https://www.acaps.org/covid19-government-measures-dataset	MEDIDAS
5	https://github.com/owid/covid-19-data/tree/master/public/data/	TESTS, GDP, CAMAS HOSPITAL; POBLACION
6	https://github.com/microsoft/Bing-COVID-19-Data/tree/master/data	ESTADISTICAS CONFIRMADOS, RECUPERDADOS FALLECIDOS

De este modo una vez definidos los datos a extraer de cada fuente se ha realizado el proceso de reestructuración y limpieza que se detalla a continuación.



Resumen datos recolectados

Clase	Cantidad instancias
Continent	**1**
Country	**14**
Province	**273**
Statistic (Country)	**1440**
Statistic (Province)	**9758**
Patient	**65633**
Test	**815**
Actions	**628**

- ## Pre-procesamiento de datos

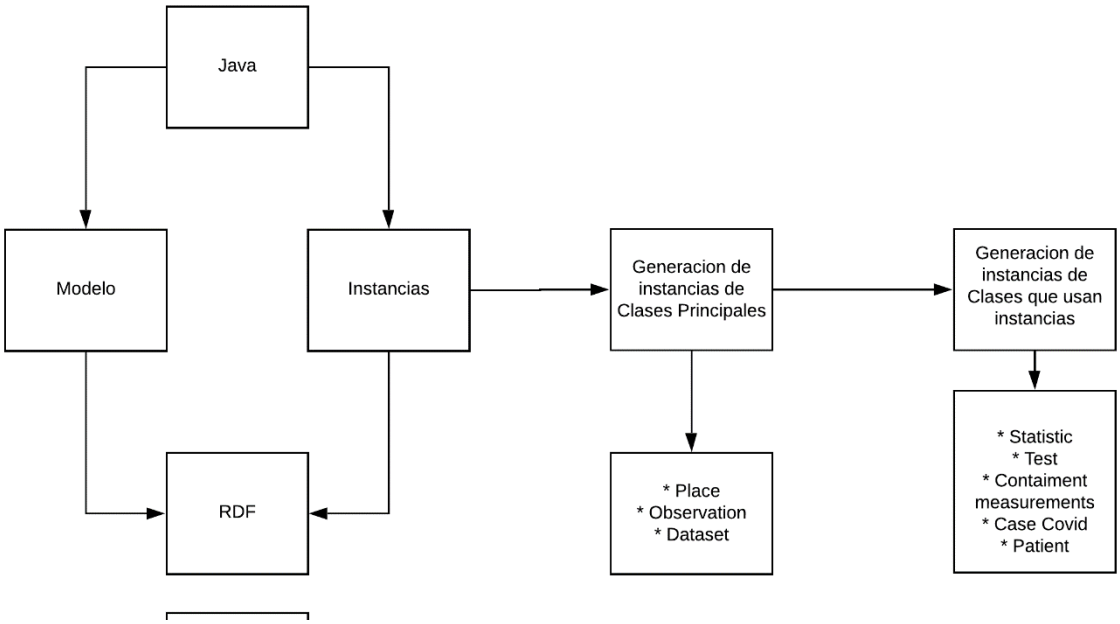
Se ha procedido a analizar métodos para adaptar los datos al modelo ontológico propuesto previo a la generación de individuos o instancias empleando Jena.

De este modo se ha considerado en función a los atributos previamente definidos clasificar las fuentes de datos en función a la data que proveen, sean estas estadísticas, medidas o acciones gubernamentales, descripción de casos, entre otras, para de este modo lograr un proceso más limpio al momento de adaptar las fuentes de datos al modelo propuesto. Así se ha definido la información a obtener de cada una de las fuentes para evitar redundancia.

Id	Fuente	Data
1	https://www.datos.gov.co/Salud-y-Protecci-n-Social/Casos-positivos-de-COVID-19-en-Colombia/gt2j-8ykr/data	INFORMACION MEDICA DE CASOS EN COLOMBIA
2	https://github.com/andrab/ecuacovid	ESTADISTICAS HOSPITALIZADOS, MUESTRAS
3	https://data.humdata.org/dataset/total-covid-19-tests-performed-by-country	TESTS
4	https://www.acaps.org/covid19-government-measures-dataset	MEDIDAS
5	https://github.com/owid/covid-19-data/tree/master/public/data/	TESTS, GDP, CAMAS HOSPITAL; POBLACION
6	https://github.com/microsoft/Bing-COVID-19-Data/tree/master/data	ESTADISTICAS CONFIRMADOS, RECUPERDADOS FALLECIDOS

De este modo una vez definidos los datos a extraer de cada fuente se ha realizado el proceso de reestructuración y limpieza que se detalla a continuación.

- # Transformación de datos



CONSULTAS

1. **¿Cuál es el número de casos, fallecidos, pruebas y recuperados por provincia?**

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX gn: <http://www.geonames.org/ontology#>
PREFIX j.0: <https://ld.utpl.edu.ec/dataCOVID/ontology#>
PREFIX schema: <http://schema.org/>
select distinct ?value ?province where {
    ?province rdf:type dbo:Province.
    ?res gn:locatedIn ?province.
    ?res j.0:quantity ?value.
    ?res schema:observationDate "31/05/2020".
} order by (?province)
```

	value	province
1	"7800"	covidData:province/AR-B
2	"5617"	covidData:province/AR-B
3	"173"	covidData:province/AR-B
4	"214"	covidData:province/AR-B
5	"0"	covidData:province/AR-B
6	"11"	covidData:province/AR-D
7	"0"	covidData:province/AR-D
8	"30"	covidData:province/AR-E
9	"0"	covidData:province/AR-E
10	"63"	covidData:province/AR-F
11	"7"	covidData:province/AR-F
12	"0"	covidData:province/AR-F
13	"22"	covidData:province/AR-G
14	"0"	covidData:province/AR-G
15	"866"	covidData:province/AR-H
16	"48"	covidData:province/AR-H
17	"0"	covidData:province/AR-H

2. ¿Cuál es la cantidad **de pruebas** que se han desarrollado por **país**?

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

PREFIX dbo: <http://dbpedia.org/ontology/>

PREFIX gn: <http://www.geonames.org/ontology#>

PREFIX j.0: <https://ld.utpl.edu.ec/dataCOVID/ontology#>

PREFIX schema: <http://schema.org/>

PREFIX ov: <http://open.vocab.org/terms/>

select ?tests ?nombrePais ?country ?dateTests ?numberTests ?population where {

    ?country rdf:type dbo:Country.

    ?country dbo:name ?nombrePais.

    ?country dbo:population ?population.

    ?tests ov:madeIn ?country.

    ?tests schema:observationDate ?dateTests.

    ?tests j.0:realized ?numberTests.

} order by (?country)
```

	tests	nombrePais	country	dateTests	numberTests	population
1	covidData:tests/realized/AR/1-5-2020	"Argentina"	covidData:country/AR	"1/5/2020"	"61530"	"45195777"
2	covidData:tests/realized/AR/1-6-2020	"Argentina"	covidData:country/AR	"1/6/2020"	"164084"	"45195777"
3	covidData:tests/realized/AR/10-4-2020	"Argentina"	covidData:country/AR	"10/4/2020"	"16379"	"45195777"
4	covidData:tests/realized/AR/10-5-2020	"Argentina"	covidData:country/AR	"10/5/2020"	"83018"	"45195777"
5	covidData:tests/realized/AR/10-6-2020	"Argentina"	covidData:country/AR	"10/6/2020"	"208519"	"45195777"
6	covidData:tests/realized/AR/11-4-2020	"Argentina"	covidData:country/AR	"11/4/2020"	"18027"	"45195777"
7	covidData:tests/realized/AR/11-5-2020	"Argentina"	covidData:country/AR	"11/5/2020"	"85158"	"45195777"
8	covidData:tests/realized/AR/11-6-2020	"Argentina"	covidData:country/AR	"11/6/2020"	"214807"	"45195777"
9	covidData:tests/realized/AR/12-5-2020	"Argentina"	covidData:country/AR	"12/5/2020"	"87547"	"45195777"
10	covidData:tests/realized/AR/12-6-2020	"Argentina"	covidData:country/AR	"12/6/2020"	"221305"	"45195777"

3. **¿Qué país tiene mayor cantidad de confirmados en Latinoamérica**?

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX gn: <http://www.geonames.org/ontology#>
PREFIX j.0: <https://ld.utpl.edu.ec/dataCOVID/ontology#>
PREFIX schema: <http://schema.org/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
select DISTINCT   ?res ?value ?date ?nombrePais ?country where {
   ?country rdf:type dbo:Country.
   ?country dbo:name ?nombrePais.
   ?res gn:locatedIn ?country.
   ?res j.0:quantity ?value.
   ?res schema:observationDate ?date.
   FILTER CONTAINS (?date, "9/6/2020")
   FILTER CONTAINS (str(?res), "/confirmed-cases/")
} ORDER BY DESC(xsd:integer(?value)) LIMIT 1
```

	res	value	date	nombrePais	country
1	covidData:statistic/confirmed-cases/BR/33208932	"1032913"	"19/6/2020"	"Brazil"	covidData:country/BR

4. **¿Qué país tiene mayor cantidad de fallecidos en Latinoamérica?**

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX gn: <http://www.geonames.org/ontology#>
PREFIX j.0: <https://ld.utpl.edu.ec/dataCOVID/ontology#>
PREFIX schema: <http://schema.org/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
select DISTINCT   ?res ?value ?date ?nombrePais ?country where {
    ?country rdf:type dbo:Country.
    ?country dbo:name ?nombrePais.
    ?res gn:locatedIn ?country.
    ?res j.0:quantity ?value.
    ?res schema:observationDate ?date.
    FILTER CONTAINS (?date, "9/6/2020")
    FILTER CONTAINS (str(?res), "/deaths-cases/")

} ORDER BY DESC(xsd:integer(?value)) LIMIT 1
```

	res	value	date	nombrePais	country
1	covidData:statistic/deaths-cases/BR/33208932	"48954"	"19/6/2020"	"Brazil"	covidData:country/BR

5. ¿Qué país tiene mayor cantidad de pacientes recuperados en Latinoamérica?

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX gn: <http://www.geonames.org/ontology#>
PREFIX j.0: <https://ld.utpl.edu.ec/dataCOVID/ontology#>
PREFIX schema: <http://schema.org/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
select DISTINCT   ?res ?value ?date ?nombrePais ?country where {
   ?country rdf:type dbo:Country.
   ?country dbo:name ?nombrePais.
   ?res gn:locatedIn ?country.
   ?res j.0:quantity ?value.
   ?res schema:observationDate ?date.
   FILTER CONTAINS (?date, "9/6/2020")
   FILTER CONTAINS (str(?res), "/recovered-cases/")
} ORDER BY DESC(xsd:integer(?value)) LIMIT 1
```

	res	value	date	nombrePais	country
1	covidData:statistic/recovered-cases/BR/33208932	"520360"	"19/6/2020"	"Brazil"	covidData:country/BR

Consultas dinámicas del sitio web

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX gn: <http://www.geonames.org/ontology#>
PREFIX j.0: <https://ld.utpl.edu.ec/dataCOVID/ontology#>
PREFIX schema: <http://schema.org/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
select DISTINCT  ?dep ?value where {
    ?dep rdf:type dbo:Province.
    ?dep dbo:name ?nombrePais.
    ?res gn:locatedIn ?dep.
    VALUES ?dep {<${idDepA}> <${idDepB}} .
    ?res j.0:quantity ?value.
    ?res schema:observationDate ?date.
    FILTER CONTAINS (str(?res), "/${dimension}/")
} GROUP BY ?province`
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX gn: <http://www.geonames.org/ontology#>
PREFIX j.0: <https://ld.utpl.edu.ec/dataCOVID/ontology#>
PREFIX schema: <http://schema.org/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
select DISTINCT  ?dep ?value where {
    ?dep rdf:type dbo:Province.
    ?dep dbo:name ?nombrePais.
    ?res gn:locatedIn ?dep.
    VALUES ?dep {<${idDepA}> <${idDepB}} .
    ?res j.0:quantity ?value.
    ?res schema:observationDate ?date.
    FILTER CONTAINS (str(?res), "/${dimension}/")
} GROUP BY ?province`
```

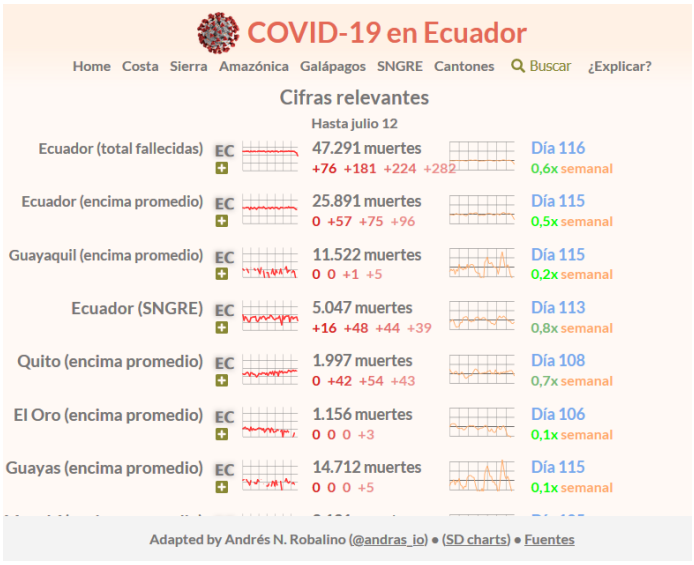
```
PREFIX qb: <http://purl.org/linked-data/cube#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

select ?dimLabel ?dep (sum(?pop) as ?popByDim) where {
?obs a qb:Observation .
        ?obs <http://id.insee.fr/meta/mesure/pop15Plus> ?pop .
        ?obs <http://id.insee.fr/meta/cog2017/dimension/DepartementOuCommuneOuArrondissementMunici
pal> ?dep .
        VALUES ?dep {<${idDepA}> <${idDepB}>} .
        ?obs <${dimension}> ?dim .
        ?dim skos:notation ?notation .
        ?dim skos:prefLabel ?dimLabel .
}
GROUP BY ?dimLabel ?dep
ORDER BY DESC(?notation)
```

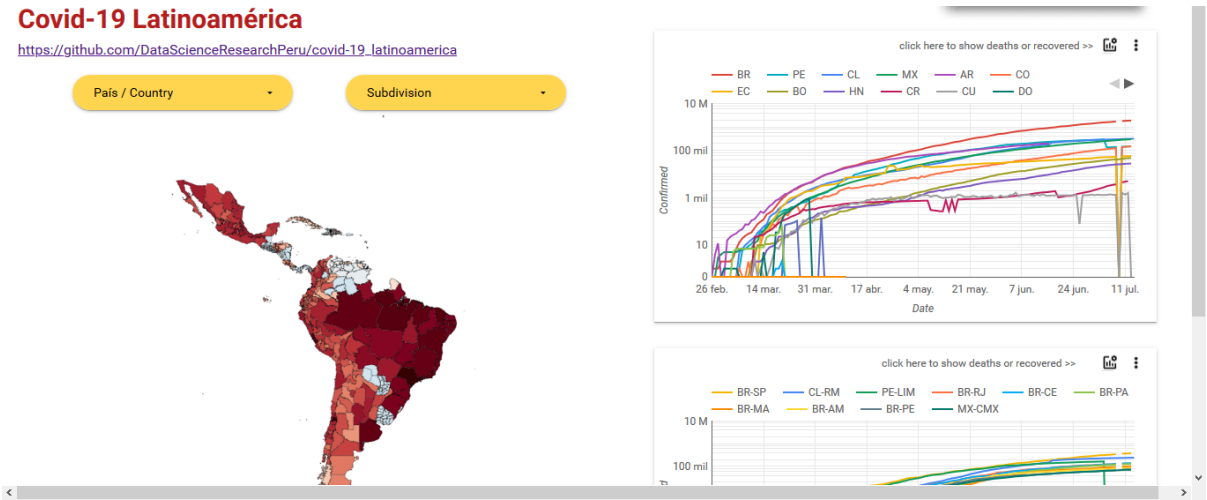
```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX gn: <http://www.geonames.org/ontology#>
PREFIX j.0: <https://ld.utpl.edu.ec/dataCOVID/ontology#>
PREFIX schema: <http://schema.org/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
select DISTINCT  ?dep ?popByDim where {
    ?dep rdf:type dbo:Province.
    ?dep dbo:name ?nombrePais.
    ?res gn:locatedIn ?dep.
    VALUES ?dep {<${idDepA}> <${idDepB}} .
    ?res j.0:quantity ?popByDim.
    ?res schema:observationDate ?date.
    FILTER CONTAINS (str(?res), "/${dimension}/")
} GROUP BY ?dep`
```

Trabajos Relacionados

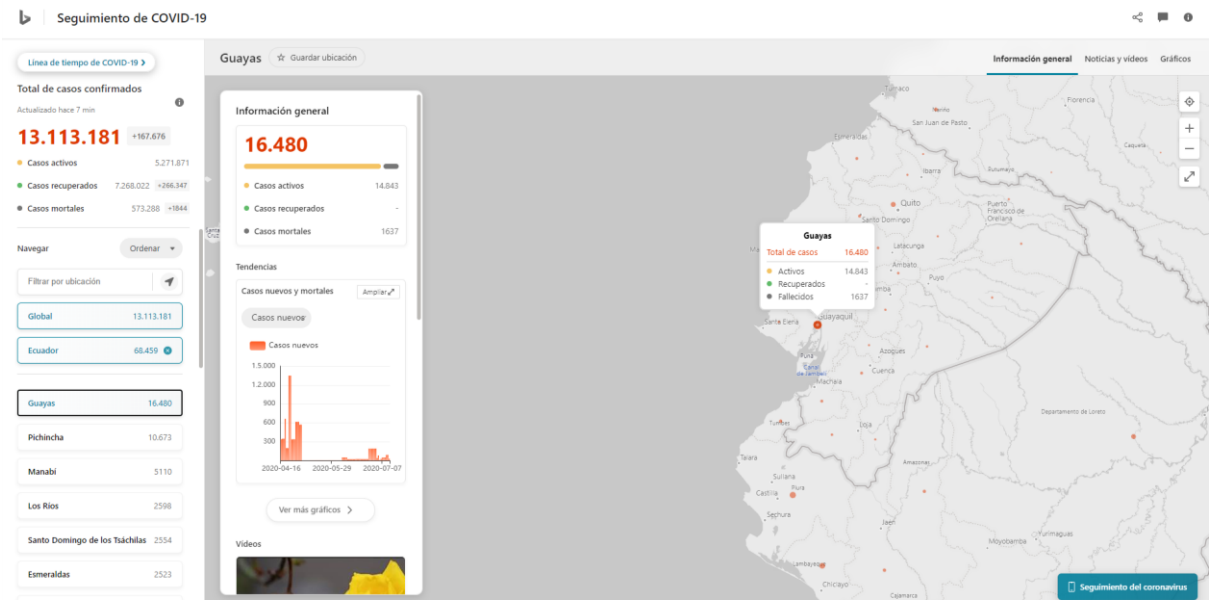
1. **Estadisticas de covid-19 Ecuador**
    a. Se trata de un sitio que presenta las estadísticas de covid en Ecuador
    b. https://covid.andresrobalino.com/



2. **Covid 19 Latin-American**
    a. Es un portal creado con la aplicación Data Studio de Google para visualización de datos que muestra las estadísticas en Sudamerica
    b. https://datastudio.google.com/u/2/reporting/9b824956-4055-46da-8c40-0d46ded5ffba/page/QkcKB



3. **Covid en el mundo**
    a. Portal creado por Bing (Microsoft) que muestra las estadísticas de covid alrededor del mundo, por país, y muestra además noticias por territorio seleccionado.
    b. https://www.bing.com/covid/

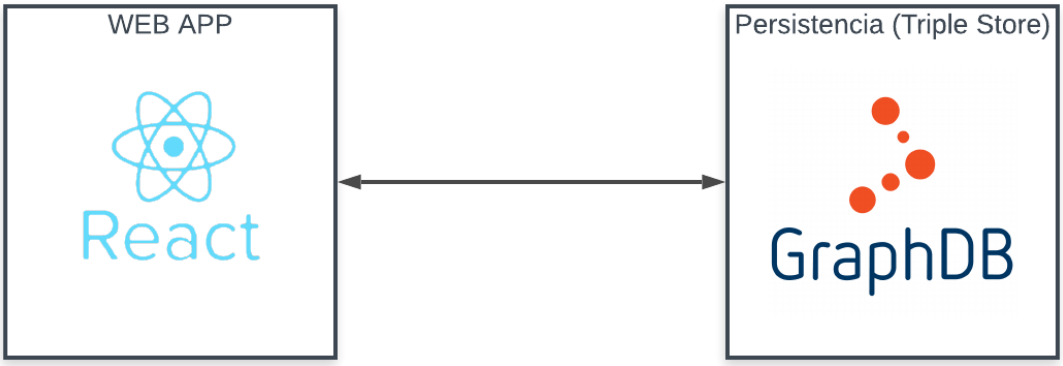


Solución propuesta

Se propone un portal mediante el cual las personas puedan realizar diversas consultas y comparaciones referentes a las estadísticas de covid en Latinoamérica.

- Comparación de estadísticas entre dos países.
- Consulta entre casos confirmados y pruebas realizadas por país
- Consulta de edades y sexo de pacientes de casos de covid descritos.
- Consulta de estadísticas de un país en fecha especifica

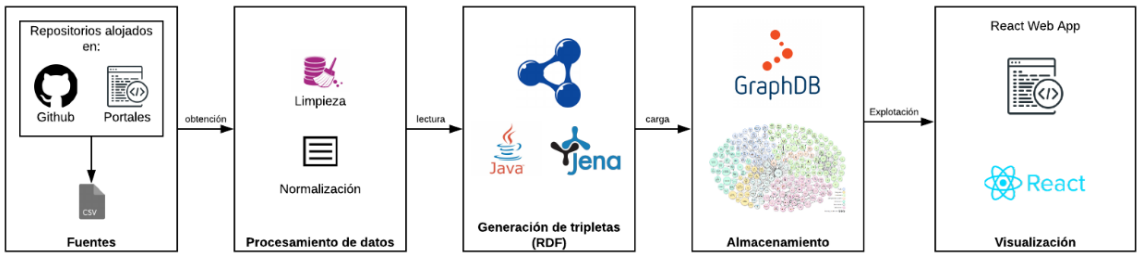
Se definirán más funcionalidades posteriormente a continuación se presenta la arquitectura en una versión inicial de la aplicación



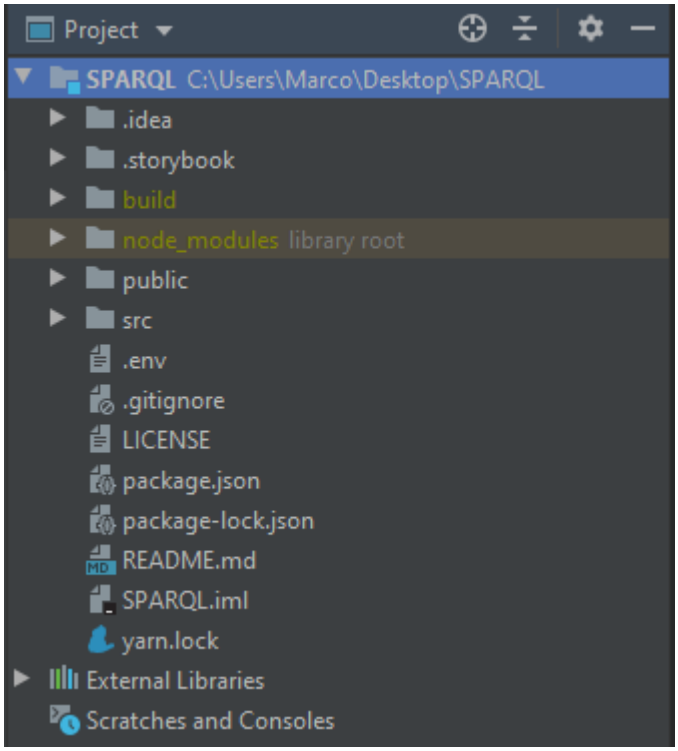
Desarrollo

Proceso

El proceso inicia con la búsqueda de fuentes de datos relacionados con estadísticas del Covid 19 en Latinoamérica, seguidamente se procede a extraer esta información para proceder a limpiarla y estructurarla como paso previo para la representación empleando el modelo ontológico, posteriormente empleando Java y Jena (librería para procesamiento LOD). Se procedió a generar las tripletas (Sujeto - Predicado - Objeto), las cuales representan la información de forma entendible para máquinas y humanos. Posteriormente las tripletas fueron alojadas en GraphDB (Triple Store) el cual provee varias ventajas, entre ellas un endpoint Sparql necesario para la explotación de datos. Finalmente se desarrolló una aplicación web empleando React Js, para de este modo permitir al usuario realizar las consultas del grafo de conocimiento sin necesidad de un conocimiento especializado.



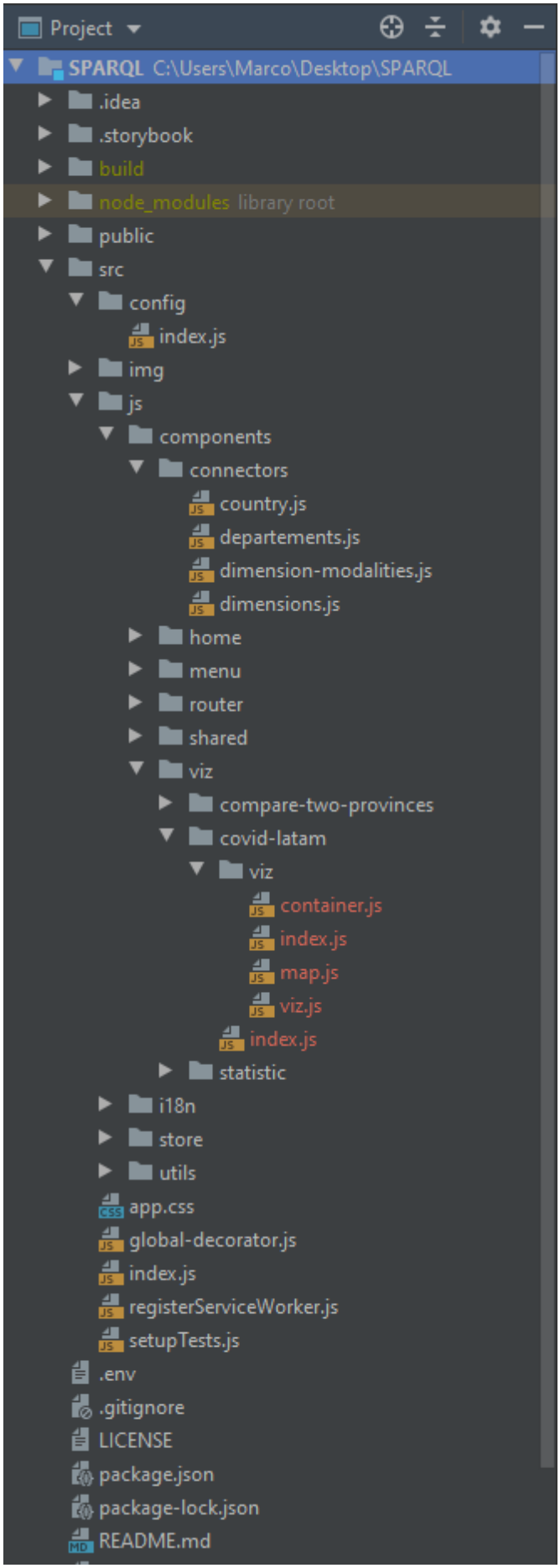
A continuación, se muestra la estructura de la aplicación React desarrollada para la explotación y visualización de datos la cual esta conformada por los siguientes ficheros



Directorio SRC: aloja el código del proyecto

package.json: Contiene las dependencias del proyecto

README.md: Contiene las instrucciones para la ejecución de la aplicación



Config/Index.js : contiene la configuración de forma global del endpoint

Connectors: contiene las sentencias SPARQL para obtener los valores para realizar las consultas dinámicas con los combos

Router: contiene la configuración de las rutas de la aplicación

Viz: contiene los componentes con las visualizaciones y formularios para que el usuario realice las consultas.

Se mantiene la misma estructura en todas las visualizaciones

container.js -> Contiene la consulta con los valores seleccionados por el usuario.

Vistas de la plataforma web