

# ***UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA***

Sistemas Informáticos y Computación

## ***Sistemas Basados en el Conocimiento***

### **Autores:**

- Edison Ruiz
- Danilo Ochoa

**Fecha:** martes, 04 de agosto del 2020.

**Data Covid-19  
NORTE Y CENTROAMÉRICA**

# Introducción

En el presente trabajo se describen las actividades llevadas a cabo para realizar el procesamiento de datos referentes a la pandemia ocurrida durante el año 2020. Los datos a traviesan por un procesamiento para ser publicados en la web semántica, proceso detallado en este documento.

Enlace a los recursos: <https://github.com/da8ah/UTPL-SBC-Covid19Project>

## Especificación de la Ontología

### 1. Documento de especificación de la ontología

Documento de Especificación de requerimientos Covid-19	
1	PROPÓSITO
	Diseñar una ontología, a partir de otras ontologías y la creación de la taxonomía necesaria, para describir un vocabulario que soporte llevar un registro de las estadísticas relacionadas a la pandemia provocada por el covid-19.
2	ALCANCE
	Definir una ontología con los metadatos recogidos de la información estadística en relación a la pandemia en lo que se refiere a número de casos categorizados desde diferentes perspectivas (p. ej. posible contagio, casos de contagios, casos de atención médica por contagio, defunciones, etc.). La ontología presente se limita a describir las estadísticas presentes en norteamérica y centroamérica únicamente.
3	LENGUAJE DE IMPLEMENTACIÓN
	OWL (Ontology Web Language) mediante la herramienta protégé.
4	POSIBLES USUARIOS FINALES
	<ol style="list-style-type: none"><li>1. Docentes y estudiantes que deseen explorar estadísticas y casos relacionados con la pandemia.</li><li>2. Agentes o sistemas, por ejemplo, predictores, bases de conocimiento, inteligencia artificial, entre otros.</li><li>3. Analistas e investigadores que necesiten datos sobre covid en norte y centroamérica.</li><li>4. Empresas públicas o privadas que requieran analizar los datos sobre Covid en norte y centroamérica.</li></ol>
5	POSIBLES USOS
	<ol style="list-style-type: none"><li>1. Consulta de información relacionada con los casos de Covid en norte y centroamérica.</li><li>2. Consulta de estadísticas relacionadas al número de casos detectados, muertes, recuperados y pruebas por país.</li><li>3. Gráficos de estadísticas</li><li>4. Publicaciones con datos abiertos</li></ol>
6	REQUERIMIENTOS
	A. REQUERIMIENTOS NO FUNCIONALES
	<ul style="list-style-type: none"><li>- La ontología debe estar en el idioma inglés.</li><li>- La ontología rehúsa ontologías y vocabularios ya existentes.</li></ul>
	B. REQUERIMIENTOS FUNCIONALES
	<ul style="list-style-type: none"><li>- ¿Cuántos casos de contagios se registraron en Nueva York en el mes de febrero?</li><li>- ¿Cuántas pruebas se realizaron en el estado (provincia) de Massachussets en el mes de febrero?</li><li>- ¿En qué estado (provincia) se presentó el mayor número de casos de contagio hasta el mes de febrero?</li></ul>

	<ul style="list-style-type: none"> <li>- ¿Cuántos casos hay registrados hasta la fecha actual globalmente?</li> <li>- ¿Cuántos de los casos registrados se recuperaron por país y por provincia (o estado) en la fecha actual dependiendo del sexo de los pacientes?</li> <li>- ¿Cuántos casos de contagios diarios ocurren por provincia (o estado) en la fecha actual en adultos mayores divididos por su sexo?</li> <li>- ¿Cuántos casos de contagios semanales ocurren por país y por provincia (o estado) en la fecha actual en un rango de edad de 25 a 65 años?</li> <li>- ¿Cuántos casos de contagios mensuales ocurren por país y por provincia (o estado) en la fecha actual dependiendo del sexo?</li> <li>- ¿Qué país y por provincia (o estado) tiene mayor cantidad de fallecidos en norte y centroamérica en la fecha actual por cada sexo?</li> <li>- ¿Qué país y por provincia (o estado) tiene mayor cantidad de pacientes recuperados en norte y centroamérica en la fecha actual por cada sexo?</li> <li>- ¿Qué país y por provincia (o estado) tiene mayor cantidad de casos detectados en norte y Centroamérica en la fecha actual por cada sexo?</li> <li>- ¿Cuántas personas de cierta nacionalidad se contagiaron en un determinado mes, semana o día?</li> </ul>	
	<b>PREGLOSARIO DE TÉRMINOS</b>	
	<ul style="list-style-type: none"> <li>- País</li> <li>- Semanal</li> <li>- Mensual</li> <li>- Contagios</li> <li>- Pruebas</li> <li>- Diarios</li> <li>- Fallecidos</li> <li>- Fuente de Datos</li> <li>- Estadísticas</li> <li>- Paciente</li> </ul>	<ul style="list-style-type: none"> <li>- Enfermedad</li> <li>- Covid</li> <li>- Estado o Provincia</li> <li>- Recuperados (Datos de alta)</li> <li>- Fecha</li> <li>- sexo</li> <li>- nacionalidad</li> <li>- fecha de nacimiento</li> </ul>

## 2. Modelo conceptual de la ontología

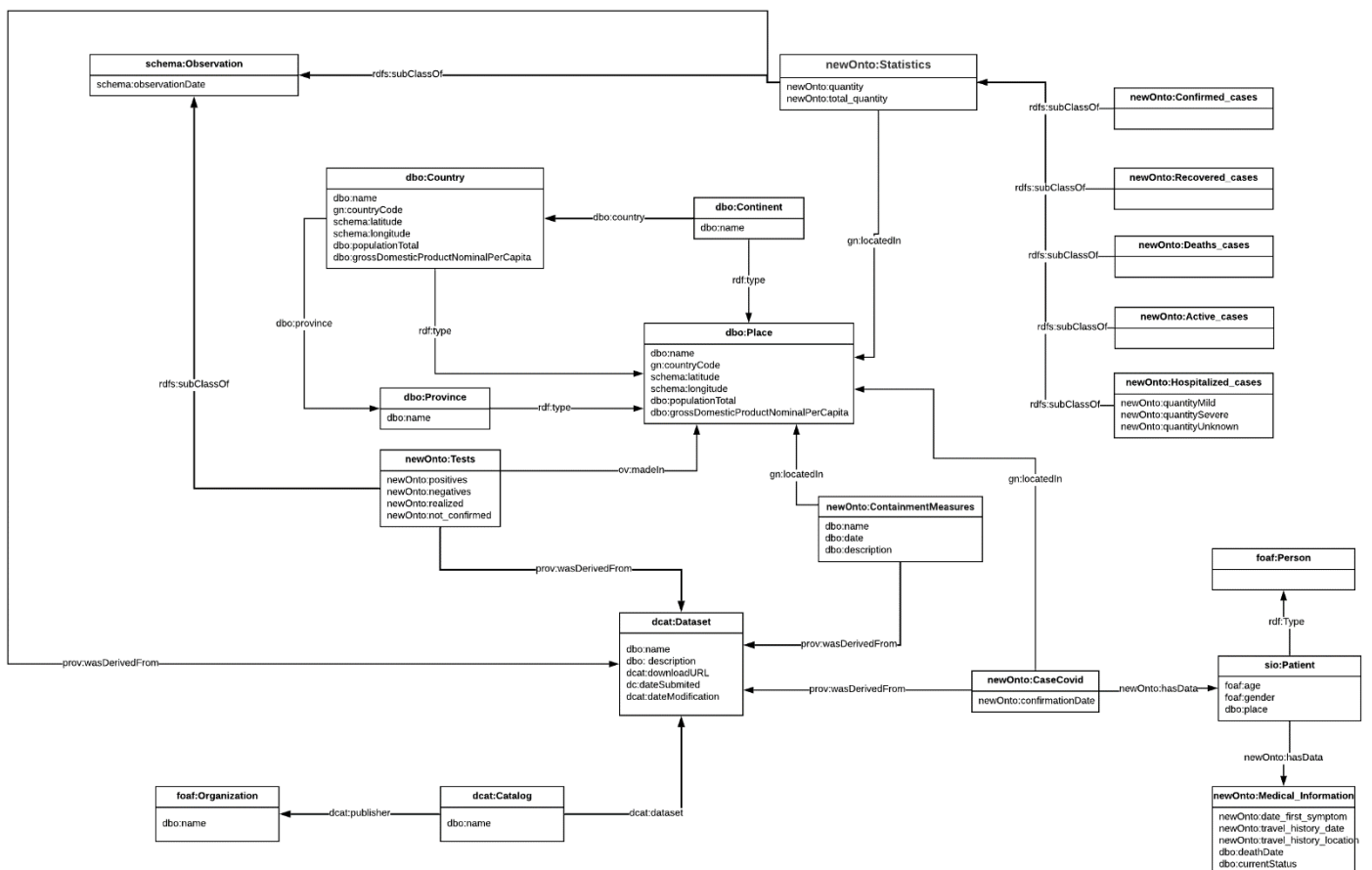
### GLOSARIO DE TÉRMINOS

NOMBRE	DESCRIPCIÓN	TIPO
Death	Cantidad de personas fallecidas por Covid-19	Concepto
Recovered	Cantidad de personas recuperadas por Covid-19	Concepto

Hospitalized	Cantidad de personas internadas en un hospital	Concepto
Test	Cantidad de pruebas Covid-19 realizadas	Concepto
Confirmed	Cantidad de personas confirmadas con Covid-19	Concepto
Continent	Nombre del continente dónde se registraron los casos de Covid-19	Concepto
Country	Nombre del país dónde se registraron los casos de Covid-19	Concepto
Province	Estado en dónde se registró el caso de Covid-19	Concepto
DataSource	Fuente de datos de donde proviene la información Acontecimiento relacionado con el Covid-19	Concepto
Statistics	Estadística del número de casos para un determinado grupo de casos	Concepto
Case	Un caso vinculado a un paciente.	Concepto
Patient	Paciente de un caso confirmado.	Concepto
updates	Actualización de datos o recursos realizada por una provincia	Propiedad
isCountryOf	Determina a qué continente corresponde un país	Propiedad

isProvinceOf	Determina a qué país corresponde una provincia (estado)	Propiedad
source	Fuente de datos de la que se obtuvo la información	Propiedad
isCompossedBy	Determina los tipos de estadísticas de una fuente	Propiedad
isRelatedWithPacient	Relaciona un caso con su paciente.	Propiedad

## Diagrama de Especificación



## Definición de URIs

Tomando como partida los lineamientos para la publicación de datos enlazados con propósitos gubernamentales para la definición de URIs se consideran los siguientes puntos [1]:

- Brindar tanta información útil sea posible mediante los URIs teniendo en cuenta que los datos están destinados a ser utilizados por ciudadanos.
- Tomar como raíz el dominio gubernamental para cada país, por ejemplo, en el caso de Estados Unidos: <http://data.usa.gov.us/>
- Diferenciar los datos de la ontología empleando, luego de la URI gubernamental, resource para los datos y ontology para la ontología, por ejemplo: <http://data.usa.gov.us/resource/>, <http://data.usa.gov.us/ontology/>
- Reutilizar vocabularios preexistentes para la ontología.
- Y por último utilizar para la definición el idioma oficial de cada país cuando sea posible, por ejemplo: <http://data.usa.gov.us/ontology/Pruebas>
- Como recomendación los URIs deben utilizar el slash, cuando sea posible, en vez del numeral para mejorar los tiempos de respuesta en conjuntos de datos enormes.

A continuación, según los lineamientos considerados, se enlistan los URIs que se proponen para este proyecto categorizándolos según el nivel de pertenencia al que corresponden, su significado y la ejemplificación del URI, tomando en cuenta lo siguiente:

- Debido a la convención entre todos los integrantes del proyecto el idioma en el que se va a construir todo es el inglés.
- Para la ontología se emplea el uso de URIs con numeral y para los datos URIs con slash.
- Por convención para diferenciar la ontología se utilizan las iniciales de cada clase en mayúscula y las iniciales de cada propiedad en minúscula.

Ontología		
Categoría	Significado	URI de ejemplo
Base	La base es el dominio raíz en donde se va a ubicar la información y de donde se van a desprender la ontología y los datos.	<a href="http://data.utpl.edu.ec/COVID19-Pandemic/">http://data.utpl.edu.ec/COVID19-Pandemic/</a>
Ontology	Dominio raíz de toda la ontología.	<a href="#">/COVID19-Pandemic/ontology#</a>
Class	Las clases de la ontología se identifican con la primera letra en mayúscula.	
prefijo:Country	Ejemplo de clase: Identifica la clase específica Country de la ontología.	<a href="#">/ontology#Country</a>
Property	Las propiedades de la ontología se identifican con la primera letra en minúscula.	

prefijo:hasData	Ejemplo de propiedad: Identifica la propiedad específica hasData.	/ontology#hasData
-----------------	-------------------------------------------------------------------------	-------------------

Data		
Categoría	Significado	URI de ejemplo
Base	La base es el dominio raíz en donde se va a ubicar la información y de donde se van a desprender la ontología y los datos.	http://data.utpl.edu.ec/COVID19-Pandemic/
Resource	Dominio raíz de todos los datos.	/COVID19-Pandemic/resource/
Country	Los países se identifican con el código ISO 3166-2	Para Estados Unidos: /resource/US
Province	Los países se identifican con el código ISO 3166-2	Para el estado de Alabama perteneciente al país de Estados Unidos: /resource/US_AL
Dataset	Cada Dataset se identifica con el nombre en mayúsculas de la página web del cuál fue tomado.	Para la fuente principal: /resource/COVID_19_OPEN_DATA
Statistic: <ul style="list-style-type: none"> <li>Confirmed</li> <li>Recovered</li> <li>Death</li> <li>Active</li> <li>Hospitalized</li> </ul>	Las estadísticas se identifican por el lugar y un identificador.	Para la estadística confirmada de Estados Unidos: /resource/confirmed_US_1 Para la estadística death en Alabama: /resource/death_US_AL_1
Test	Los tests se identifican por el lugar y un identificador.	Para el test de Estados Unidos: /resource/test_US_1 Para el test en Alabama: /resource/test_US_AL_1
Containment Measures	Las medidas de contención se identifican por un número único generado en secuencia.	Para la primera medida de contención en Estados Unidos: /resource/containment_measure_US_1 Para el test en Alabama: /resource/containment_measure_US_AL_1

## Licencias

La licencia que se empleará es de tipo Creative Commons (CC BY-NC-SA) la cual permite a otros remezclar, adaptar y construir sobre su trabajo de manera no comercial, siempre que lo acrediten y otorguen licencias de sus nuevas creaciones bajo los mismos términos.

Esta licencia se ha seleccionado debido a que los datos son de uso público para fines de investigación y por la situación se considera información que debe compartirse libremente. En adición, permite la posibilidad a los demás de reutilizar los modelos desarrollados en este proyecto bajo términos de libre uso.

## FUENTES DE DATOS

### Fuentes De Datos Covid-19

Nombre de la Fuente de Datos: La Prensa

Sitio Web: <https://github.com/GoogleCloudPlatform/covid-19-open-data>

Formato del archivo: CSV

### Fuentes De Datos de Medidas Gubernamentales

Nombre de la Fuente de Datos: COVID19 GOVERNMENT MEASURES DATASET

Sitio Web: <https://www.acaps.org/covid19-government-measures-dataset>

Formato del archivo: CSV

## Limpieza de Datos

Para constatar que los datos sean consistentes de acuerdo con el modelo se estructuraron en base a este. Las tareas de limpieza realizadas fueron:

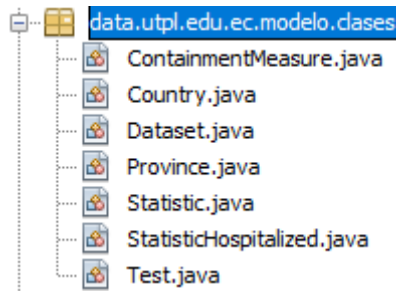
1. El identificador de países y provincias se propuso de acuerdo al estándar ISO 3166-2.
2. Los datos estadísticos (statistics) y de pruebas (tests) en formato numérico enteros no negativos.
3. Las fechas se definieron siguiendo el patrón yyyy-mm-dd (año, mes, día en la forma 2020-01-31).

Los datos se limpiaron de forma manual en los archivos CSV dejando varios archivos los cuales se usaron luego del procesamiento para generar datos en formato rdf.

## Generación RDF

Para generar los datos RDF se tomaron los datos procesados y de acuerdo con estos se crearon clases que representasen el modelo previamente definido para mapear los datos:





Seguidamente, se definió una clase que representase el modelo. Esta clase cuenta con los métodos necesarios para agregar uno por uno los datos al modelo. Finalmente, también permite imprimir tanto el modelo ontológico como los datos agregados en el modelo.

```
public ModeloCovid19() {
    // Prefijos
    model.setNsPrefix("utplOnto", utplOntoPrefix);
    model.setNsPrefix("utplData", utplDataPrefix);

    // Prefijos Externos
    model.setNsPrefix("schema", schema);
    model.setNsPrefix("dbo", dbo);
    model.setNsPrefix("dbr", dbr);
    model.setNsPrefix("dbp", dbp);
    model.setNsPrefix("gn", gn);
    model.setNsPrefix("dcat", dcat);
    model.setNsPrefix("prov", prov);
    model.setNsPrefix("dcterms", DCTerms.getURI());
    crearModeloOntologico();
}
```

Finalmente, desde los archivos resultantes del procesamiento para la limpieza, los cuales permanecieron en formato CSV, se realizó la lectura en la clase GeneradorRDF para luego enviar los datos al modelo e imprimirlos.

```
public static void main(String[] args) {
    // TODO code application logic here

    // Las funciones deben ser llamadas en un orden específico:
    modeloCovid19.imprimirModeloOntologico();
    cargarCountriesDesdeArchivo();
    cargarProvincesDesdeArchivo();
    cargarDatasetsDesdeArchivo();
    cargarMainDataSource();
    cargarContainmentMeasuresDesdeArchivo();
    modeloCovid19.imprimirModelo();
}
```


Como se mencionó anteriormente se generó un archivo en formato rdf para el modelo ontológico y otro archivo en formato rdf para los datos.

## Subida de datos rdf al repositorio GraphDB

Los datos rdf deben ser almacenados en un repositorio de datos semánticos por lo que se seleccionó GraphDB como tecnología de almacenamiento debido a que su velocidad de búsqueda depende únicamente del número de relaciones concretas, no del conjunto de datos, añadiendo que sus estructuras son flexibles y ágiles, siendo muy eficiente y presentando resultados en tiempo real.

Class	Links	
dcat:Dataset	495K	← -
dbo:Province	495K	⇔ -
utplOnto:Death	456K	→ -
utplOnto:Confirmed	420K	→ -
utplOnto:Test	46K	→ -
utplOnto:Recovered	42K	→ -
utplOnto:Hospitalized	26K	→ -
dbo:Country	875	← -
utplOnto:Containment Measure	576	→ -

Local

 **utplCovid19 • UTPL COVID-19**

total statements  
**3,011,663**

3,011,663 explicit  
0 inferred  
1.00 expansion ratio

[Import RDF data](#)

[Import tabular data with OntoRefine](#)

[Export RDF data](#)

## Consultas realizadas al repositorio GraphDB

Para realizar consultas se emplea el lenguaje SPARQL. A continuación se observan algunos de los recursos obtenidos resultantes de consultar el todos los casos confirmados en adición con la fecha del suceso devolviendo la consulta el recurso en la columna izquierda y la fecha en la columna derecha.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX utplData: <http://data.utpl.edu.ec/COVID19-Pandemic/resource/>
PREFIX gn: <http://www.geonames.org/ontology#>
PREFIX utplOnto: <http://data.utpl.edu.ec/COVID19-Pandemic/ontology#>
PREFIX schema: <http://schema.org/>
SELECT ?statistic ?date
WHERE {
    ?statistic rdf:type utplOnto:Confirmed ;
    schema:observationDate ?date .
} ORDER BY ?date
```

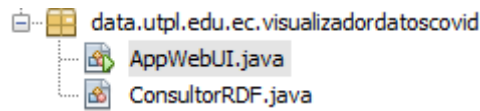
Filter query results		Showing results from 1 to 1,000 of 209,963. Query took 9.4s, today at 02:43.	
	statistic		date
1	utplData:confirmed_US_1		"2020-01-01"
2	utplData:confirmed_US_2		"2020-01-02"
3	utplData:confirmed_US_3		"2020-01-03"
4	utplData:confirmed_US_4		"2020-01-04"
5	utplData:confirmed_US_5		"2020-01-05"
6	utplData:confirmed_US_6		"2020-01-06"
7	utplData:confirmed_US_7		"2020-01-07"
8	utplData:confirmed_US_8		"2020-01-08"
9	utplData:confirmed_US_9		"2020-01-09"
10	utplData:confirmed_US_10		"2020-01-10"

## Aplicación desarrollada para el consumo de los datos

La aplicación realiza una consulta al repositorio de GraphDB y presenta los resultados en una tabla. Para realizar esto se emplearon las siguientes tecnologías:

1. Java (lenguaje de programación).
2. SparkJava (framework para desarrollo de aplicaciones web).
3. GraphDB para almacenar los datos RDF.
4. RDF4J es una librería que permite conectar con un repositorio rdf para obtener los datos resultantes de una consulta.

El proyecto consta de dos partes, un visualizador y un consultor. El consultor conecta con el repositorio y el visualizador despliega los datos en una página web.



## Recursos del mes de febrero en United States.

N	Recurso	Fecha
1	<a href="http://data.utpl.edu.ec/COVID19-Pandemic/resource/confirmed_US_32">http://data.utpl.edu.ec/COVID19-Pandemic/resource/confirmed_US_32</a>	2020-02-01
2	<a href="http://data.utpl.edu.ec/COVID19-Pandemic/resource/death_US_32">http://data.utpl.edu.ec/COVID19-Pandemic/resource/death_US_32</a>	2020-02-01
3	<a href="http://data.utpl.edu.ec/COVID19-Pandemic/resource/confirmed_US_33">http://data.utpl.edu.ec/COVID19-Pandemic/resource/confirmed_US_33</a>	2020-02-02
4	<a href="http://data.utpl.edu.ec/COVID19-Pandemic/resource/death_US_33">http://data.utpl.edu.ec/COVID19-Pandemic/resource/death_US_33</a>	2020-02-02
5	<a href="http://data.utpl.edu.ec/COVID19-Pandemic/resource/confirmed_US_34">http://data.utpl.edu.ec/COVID19-Pandemic/resource/confirmed_US_34</a>	2020-02-03
6	<a href="http://data.utpl.edu.ec/COVID19-Pandemic/resource/death_US_34">http://data.utpl.edu.ec/COVID19-Pandemic/resource/death_US_34</a>	2020-02-03
7	<a href="http://data.utpl.edu.ec/COVID19-Pandemic/resource/confirmed_US_35">http://data.utpl.edu.ec/COVID19-Pandemic/resource/confirmed_US_35</a>	2020-02-04
8	<a href="http://data.utpl.edu.ec/COVID19-Pandemic/resource/death_US_35">http://data.utpl.edu.ec/COVID19-Pandemic/resource/death_US_35</a>	2020-02-04
9	<a href="http://data.utpl.edu.ec/COVID19-Pandemic/resource/confirmed_US_36">http://data.utpl.edu.ec/COVID19-Pandemic/resource/confirmed_US_36</a>	2020-02-05
10	<a href="http://data.utpl.edu.ec/COVID19-Pandemic/resource/death_US_36">http://data.utpl.edu.ec/COVID19-Pandemic/resource/death_US_36</a>	2020-02-05
11	<a href="http://data.utpl.edu.ec/COVID19-Pandemic/resource/confirmed_US_37">http://data.utpl.edu.ec/COVID19-Pandemic/resource/confirmed_US_37</a>	2020-02-06
12	<a href="http://data.utpl.edu.ec/COVID19-Pandemic/resource/death_US_37">http://data.utpl.edu.ec/COVID19-Pandemic/resource/death_US_37</a>	2020-02-06

## Conclusiones

- El desarrollo de este proyecto permitió entender el proceso que se debe realizar para la publicación de datos en la web semántica.
- Existen diferentes formas para el tratamiento de los datos, para procesarlos, realizar una limpieza y obtener un resultado satisfactorio.
- La aplicación de estándares es importante para tener una forma generalizada de entendimiento y que el trabajo desarrollado pueda ser útil para otros investigadores.