UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA

Sistemas Basados en Conocimientos

Tema: Informe de datos RDF

Estudiantes: Alexis Montoya

Juan Ramón

Luis Rojas

Definición de URIs y preparación de datos origen. Indicar qué criterios consideraron para asignar o generar las URIs de los diferentes recursos.

Para definir las URIs se consideró el tipo de documento para generar RDF, ya que se tratan de datos científicos se usa las siguientes URIs:

- bibo:Document
- bibo:Journal
- bibo:Book
- foaf:Person

Transformación y almacenamiento de datos RDF. Indicar:

Tabla resumen de datos recolectados: Por cada clase del modelo ontológico indicar cuántas instancias generaron.

Clase	Total-Instancias
bibo:Document	**534**
foaf:Person	**921**
bibo:Book	**214**
Bibo:Journal	**309**

Preprocesamiento de datos: Indicar qué tareas de limpieza o transformación de datos realizaron antes de generar RDF.

- **Para la limpieza de datos se realiza lo siguiente:**
  - Primeramente, mediante el uso de la Api de Crosref se extrae toda la data a usar. Luego mediante la Api de semantic scholar se valida si hay datos extras por agregar.

  - Para la limpieza de los datos se borra los espacios en blanco del Excel generado. Una vez borrado los espacios en blancos se procede a darle formato a la columna fecha (dd-mm-aaaa), ya que esta trae otro formato.

- Se unen todas las columnas de autores (author1, author2, author3).

- **Para la transformación de datos se realiza lo siguiente:**

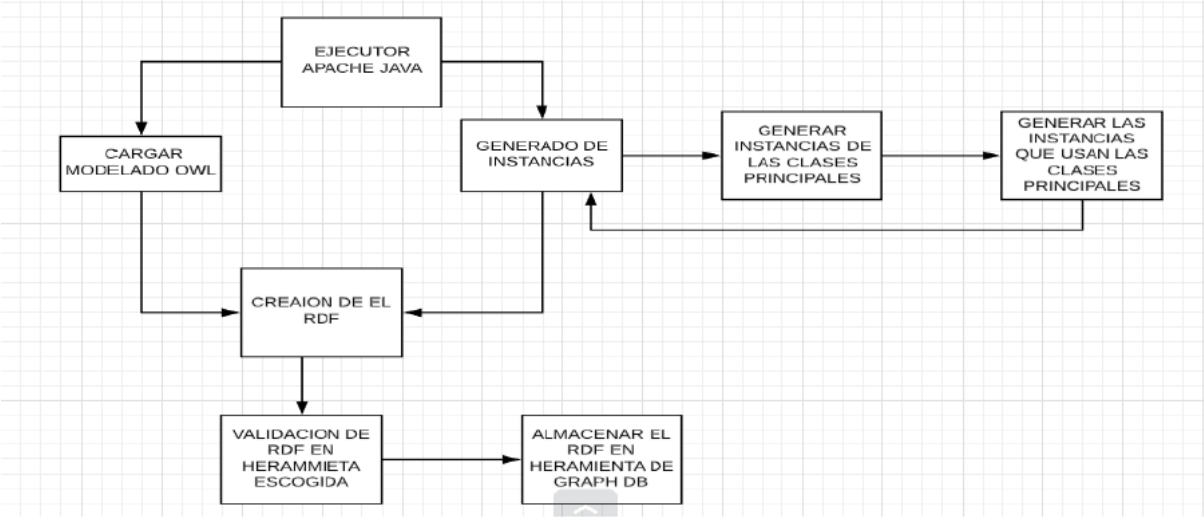
  - Primeramente, he procedido a analizar métodos para adaptar los datos al modelo ontológico propuesto previo a la generación de individuos o instancias empleando Jena. Para ello lo que se ha hecho es definir las fuentes principales de las cuales se va a sacar los datos, en función de los atributos que ya fueron definidos con anterioridad.

  - Se ha clasificado de un dataset los atributos que solamente constan en el modelo, eliminando así los atributos innecesarios o con poca funcionalidad para nuestro modelo, con esto se pudo conseguir una limpieza más optima de datos para poderlos adaptar al modelo propuesto y para observar de una mayor manera dichos datos, se analizó cada dato de forma manual, esto con la finalidad para que se pueda evitar redundancia en los datos.

Transformación de datos: Indicar la lógica del motor de transformación de datos basado en Jena. En este punto se puede indicar algún esquema que resuma los métodos y demás objetos que se implementaron.

Una vez escogidas las fuentes a trabajar se estará empleando el siguiente modelo de proceso, para la construcción, modelado y limpieza.

Este sería el proceso que se usaría, para finalizar toda la transformación:



Almacenamiento: indicar cuál fue el repositorio utilizado para almacenar los datos y las razones para realizar tal elección.

Se uso GraphDB, ya que permite almacenar, organizar y gestionar los datos de forma semántica, además de que permite la compatibilidad con datos abiertos.

Enlazado post-transformación: incluir cuántas tripletas adicionales se encontraron en DBPedia y qué metadatos se anotaron (podría ser extendido).

Para poder realizar este paso modificamos el código (**revisar imagen 1**). Dentro del archivo file.txt se encuentran todos los doi de nuestro documento y en base a eso ira realizando la búsqueda en dbpedia.

Se adjunta captura de pantalla de **file.txt** y **text_3.txt.**

Los archivos generados se encuentran en el GitHub del proyecto.

Imagen 1

```
# To get annotations in a iterative way from the source:
#paper_id = 'http://dx.doi.org/10.14213/inteuniorigh.27.3.0020' # id or some identifier
#text = """We propose a fact validation challenge for ISWC 2019. The participants will be provided set of facts in the form o
#dbCategories = getAnnotations(paper_id, text)
with open("file.txt", "r") as tf:
    lines = tf.read().split(',')
    for line in lines:
        # To get annotations in a iterative way from the source:
        paper_id = line # id or some identifier
        text = """We propose a fact validation challenge for ISWC 2019. The participants will be provided set of facts in the
        dbCategories = getAnnotations(paper_id, text)
        fic = open("text_3.txt", "a+")
        fic.writelines("%s\n" % s for s in dbCategories)

fic.close()
```

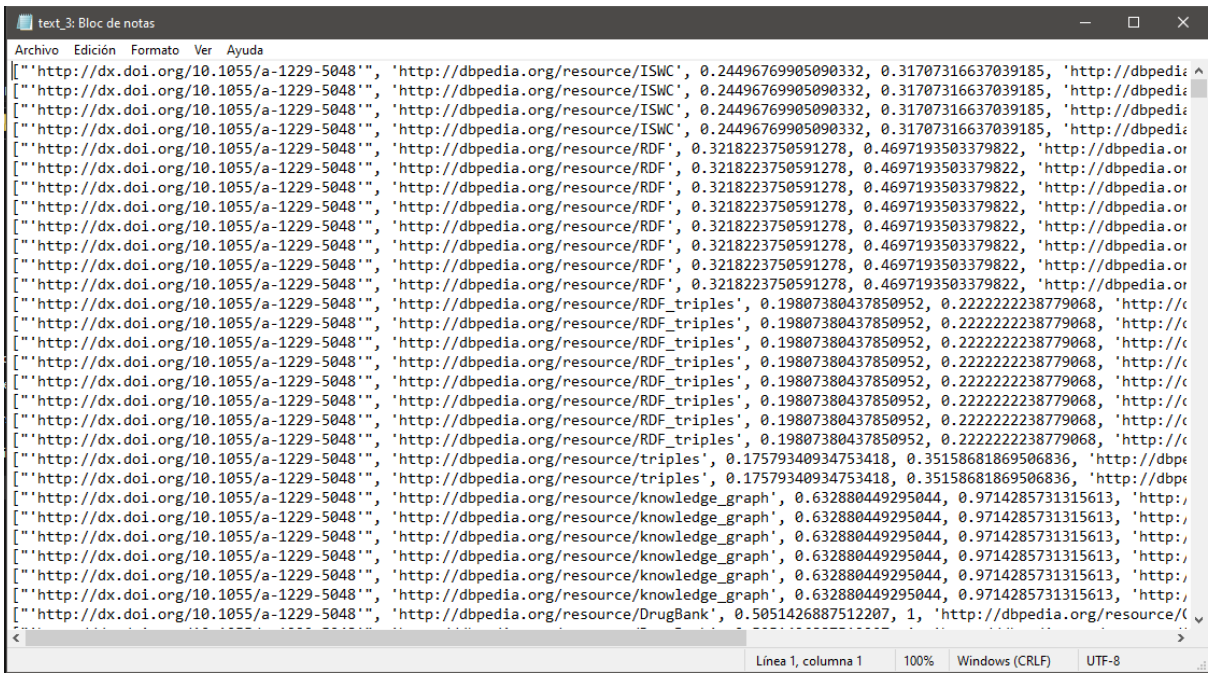
file.txt

file: Bloc de notas

Archivo Edición Formato Ver Ayuda

'http://dx.doi.org/10.1055/a-1229-5048',
'http://dx.doi.org/10.26524/royal.37.21',
'http://dx.doi.org/10.26524/royal.37.15',
'http://dx.doi.org/10.26524/royal.37.25',
'http://dx.doi.org/10.26524/royal.3718',
'http://dx.doi.org/10.26524/royal.37.18',
'http://dx.doi.org/10.26524/royal.37.17',
'http://dx.doi.org/10.26524/royal.37.27',
'http://dx.doi.org/10.26524/royal.37.20',
'http://dx.doi.org/10.26524/royal.37.26',
'http://dx.doi.org/10.26524/royal.37.22',
'http://dx.doi.org/10.26524/royal.37.9',
'http://dx.doi.org/10.26524/royal.37.16',
'http://dx.doi.org/10.26524/royal.37.5',
'http://dx.doi.org/10.26524/royal.37.19',
'http://dx.doi.org/10.26524/royal.37.28',
'http://dx.doi.org/10.1158/1557-3265.covid-19-ia26',
'http://dx.doi.org/10.26524/royal.37.13',
'http://dx.doi.org/10.26524/royal.37.29',
'http://dx.doi.org/10.26524/royal.37.6',
'http://dx.doi.org/10.5603/fc.2021.0010',
'http://dx.doi.org/10.26524/royal.37.4',
'http://dx.doi.org/10.23880/oajprs-16000135',
'http://dx.doi.org/10.26524/royal.37.23',
'http://dx.doi.org/10.1093/law-occ19/e15.013.15',
'http://dx.doi.org/10.1093/law-occ19/e4.013.4',
'http://dx.doi.org/10.1093/law-occ19/e5.013.5',
'http://dx.doi.org/10.1093/law-occ19/e12.013.12',
'http://dx.doi.org/10.1093/law-occ19/e7.013.7',

text_3



Enlace Github (código + rdf + instancias generadas)

https://github.com/Jtaramon/ProyectoSbc/tree/main/Codigo