OBTENCIÓN DE DATOS

Práctica integral de la asignatura
Curso 2017-2018

OBJETIVO DE LA PRÁCTICA

Este curso académico la práctica de la asignatura *Obtención de Datos* tiene como objetivo la generación de un dataset en formato RDF/XML sobre el transporte público en Londres, a partir de la obtención de diversos datasets y de la integración de los mismos.

CONTEXTO DE LA PRÁCTICA

Transport for London (TfL) es el organismo del gobierno local responsable de la mayoría de los aspectos del sistema de transportes en Londres [1]. En relación al desarrollo de esta práctica concreta es conveniente conocer que TfL contempla, entre otros, el transporte en:

- Metro (London Underground –Tube-, https://en.wikipedia.org/wiki/London_Underground
- Tren suburbano (London Overground, https://es.wikipedia.org/wiki/London_Overground);
- Metro ligero (Docklands Light Railway –DLR-, https://en.wikipedia.org/wiki/Docklands_Light_Railway).

TfL sigue la tendencia actual de política de datos abiertos. Por ello ofrece la posibilidad de descargar sus datasets de transporte público a través de APIs [2]. La documentación acerca de los data feeds está disponibles en [3].

REALIZACIÓN DE LA PRÁCTICA

La realización de la práctica se llevará a cabo en grupos. Los grupos estarán formados por 2 ó 3 alumnos de forma obligatoria.

La práctica consiste en el desarrollo de la descarga e integración de dos datasets de datos de transporte público, en formato XML, y de la generación de un único dataset en formato RDF/XML. El desarrollo se llevará a cabo con el lenguaje de programación Python.

Los dos datasets origen son:

a) Station facilites: Our station facilities feed is a Geo-coded KML feed of most London Underground, DLR and London Overground stations. It has station facilities and access information for each station [4].
b) Step free tube guide data: The data contained in this feed provides information about the level of step-free access to platforms and trains that is available at London Underground, London Overground and DLR stations [5].

El proceso y los pasos obligatorios a seguir para el desarrollo de la práctica son los siguientes:

1) Acceder a la página de datos abiertos de TfL y descargar ambos datasets a través de API.
2) Generar un fichero similar a a), denominado *StationFacilitiesNOH*, donde no se incorpore la información asociada a las etiquetas `<openingHours>xxxx</openingHours>`, ni las propias etiquetas .
3) Generar un fichero similar a b), denominado *StepFreeTubeNNone*, que no debe incorporar la información asociada a las etiquetas `<AccessibilityType>None</AccessibilityType>` (ni las propias etiquetas) cuando el contenido sea `None`.
4) Elaborar un diagrama conceptual con las entidades y las relaciones entre dichas entidades, que soporte la información del dominio de este problema, es decir, teniendo en cuenta las entidades de los ficheros StationFacilitiesNOH y de StepFreeTubeNNone. Para el diagrama conceptual ver la transparencia número 76 del tema *Datos Semánticos y Enlazados*.
5) Generar un fichero XML denominado *TFLfacilities,* resultado de la integración de los ficheros del apartado 2 y 3.
6) Generar un grafo RDF (que será el resultado de la integración de los ficheros antes mencionados), utilizando la librería *rdflib*. Para ello sería conveniente seguir el diagrama conceptual creado en el punto 4. El resultado de este último paso será un fichero RDF/XML. Hay que tener en cuenta que la información deberá integrarse asociando los nombres de las estaciones de a) y b) cuando estos sean iguales. A continuación se muestra un ejemplo para el caso concreto de una estación.

La estación *Beckton* en a) aparece como:

```
<station xmlns="" id="1002011" type="dlr">
        <name>Beckton</name>
…
```

Y la estación *Beckton* en (b) aparece como:

```
<Station>
        <StationName>Beckton</StationName>
…
```

Por lo tanto, como los datos asociados al elemento `<name>` del fichero XML de a) y de `<StationName>` del fichero XML de b) son iguales (en este caso concreto son iguales a *Beckton*), deberán integrarse los elementos de `<station xmlns="" id="1002011" type="dlr">` y de `<Station>` de ambos ficheros para generar el rdf.

Sobre el vocabulario a utilizar, puede definirse el espacio de nombres siguiente: `xmlns:tfl="http://tfl.gov.uk/tfl#"`. Esto permitiría que en el fichero RDF/XML los datos aparecieran nombrados así, por ejemplo: `<tfl:name>Beckton</tfl:name>` o `<tfl:Station>`.

ENTREGA DE LA PRÁCTICA

MATERIAL A ENTREGAR

A. Un ZIP denominado "Datos_*ApellidosNombre*" (Apellidos y nombre será de sólo uno de los alumnos del grupo de prácticas) con los ficheros que dan lugar a los ficheros de datos y el diagrama conceptual de los pasos 1, 2, 3, 4, 5 y 6.
B. Otro ZIP denominado "Codigo_*ApellidosNombre*" (Apellidos y nombre será de sólo uno de los alumnos del grupo de prácticas) con los ficheros de código necesarios para llevar a cabo los pasos 1, 2, 3, 5 y 6. **Hay que entregar los ficheros .py con el código fuente, no un notebook. Preparar el fichero comprimido que entreguéis con los ficheros de datos necesarios del punto A, para que al ejecutar el código dichos ficheros estén disponibles y no haya que copiarlos en vuestro directorio para la ejecución.**

La entrega se realizará a través del Aula Virtual, en dos enlaces disponibles para ello (la entrega de A y la entrega de B) tanto en la convocatoria ordinaria (como en la extraordinaria para aquellos alumnos que lo necesitaran).

FECHA DE ENTREGA DE LA PRÁCTICA

La fecha límite para la entrega de la práctica en la convocatoria ordinaria será el **día 17 de enero de 2018** a las 9:00 AM.

Para la convocatoria extraordinaria, la fecha límite será el día 16 de julio de 2018 a las 9:00 AM.

Se desea resaltar que, si algún grupo de alumnos tuviese problemas de tiempo, se podrá realizar la entrega en cualquier momento posterior al día 17 de enero. En dicho caso la entrega se realizaría a través de los enlaces de la convocatoria extraordinaria ya que dicha entrega tendrá la consideración de evaluación extraordinaria. Por lo tanto, el resultado de la evaluación de estas será con fecha posterior al día 16 de julio.

ENLACES DE INTERÉS PARA LA REALIZACIÓN DE LA PRÁCTICA

[1]. TFL (Transport for London), Wikipedia, https://en.wikipedia.org/wiki/Transport_for_London
[2]. Página de datos abiertos de TFL (Transport for London Unified API), https://api.tfl.gov.uk/
[3]. Documentación acerca de los data feeds de TFL, https://api-portal.tfl.gov.uk/docs
[4]. Station Facilities, https://data.tfl.gov.uk/tfl/syndication/feeds/stations-facilities.xml?app_id=3a0c4a01&app_key=ea95c6674605181c5dde64c7bb5d883c
[5]. Step Free Tube Guide, https://tfl.gov.uk/tfl/syndication/feeds/step-free-tube-guide.xml