1.- Introducción

Este trabajo considera tomar un conjunto de datos en formato CSV y transformarlos para dejarlos disponibles para la web semántica.

La idea es trabajar con las farmacias existentes en el ayuntamiento de Gijón.

2.- Proceso de Transformación

a.- Selección de la fuente de datos, donde se explique el conjunto de datos que se ha seleccionado para transformar, especificando su origen

Los datos corresponden a las farmacias del ayuntamiento de Gijón, estos datos están accesibles a través de un archivo CSV.

La url desde donde se obtiene el catálogo CSV:

http://datos.gob.es/es/catalogo/l01330241-farmacias

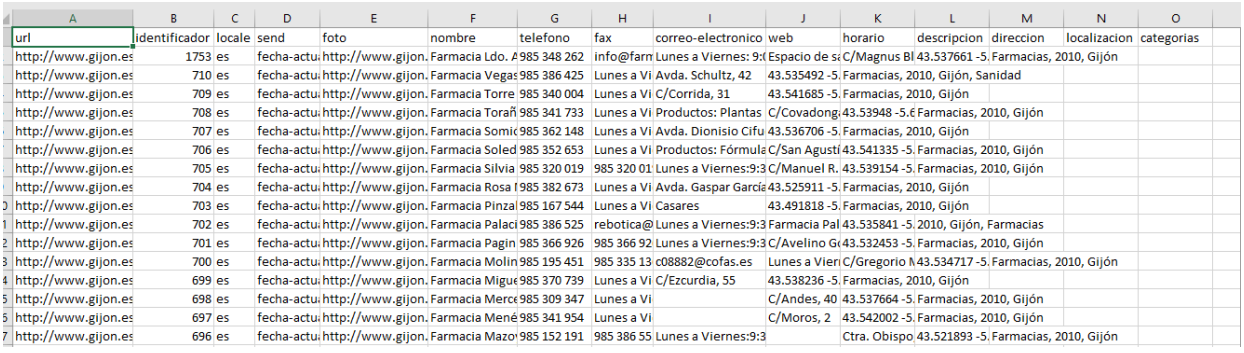
Estos datos están bajo la licencia:

http://creativecommons.org/licenses/by/3.0/es

b.- Análisis de los datos, explicando que tipo de datos se manejan, su formato, tipos de valores, y en general cualquier aspecto relevante para su transformación y explotación. Este análisis debe incluir la licencia de origen de los datos y la justificación de la licencia a usar en los datos transformados

Los datos relacionados con las farmacias están disponibles en formato CSV, para su análisis inicial se utiliza Excel.

La información cuenta con 104 registros, donde cada uno de estos corresponde a una farmacia, en la siguiente figura aparece una muestra de los datos.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	url	identificador	locale	send	foto	nombre	telefono	fax	correo-electronico	web	horario	descripcion	direccion	localizacion	categorias
	http://www.gijon.es	1753	es	fecha-actu	http://www.gijon.	Farmacia Ldo. A	985 348 262	info@farm	Lunes a Viernes: 9:(	Espacio de sa	C/Magnus Bl	43.537661 -5.	Farmacias, 2010, Gijón		
	http://www.gijon.es	710	es	fecha-actu	http://www.gijon.	Farmacia Vegas	985 386 425	Lunes a Vi	Avda. Schultz, 42	43.535492 -5.	Farmacias, 2010, Gijón, Sanidad				
	http://www.gijon.es	709	es	fecha-actu	http://www.gijon.	Farmacia Torre	985 340 004	Lunes a Vi	C/Corrida, 31	43.541685 -5.	Farmacias, 2010, Gijón				
	http://www.gijon.es	708	es	fecha-actu	http://www.gijon.	Farmacia Torañ	985 341 733	Lunes a Vi	Productos: Plantas	C/Covadong	43.53948 -5.6	Farmacias, 2010, Gijón			
	http://www.gijon.es	707	es	fecha-actu	http://www.gijon.	Farmacia Somi	985 362 148	Lunes a Vi	Avda. Dionisio Cifu	43.536706 -5.	Farmacias, 2010, Gijón				
	http://www.gijon.es	706	es	fecha-actu	http://www.gijon.	Farmacia Soled	985 352 653	Lunes a Vi	Productos: Fórmula	C/San Agustí	43.541335 -5.	Farmacias, 2010, Gijón			
	http://www.gijon.es	705	es	fecha-actu	http://www.gijon.	Farmacia Silvia	985 320 019	985 320 01	Lunes a Viernes:9:3	C/Manuel R.	43.539154 -5.	Farmacias, 2010, Gijón			
	http://www.gijon.es	704	es	fecha-actu	http://www.gijon.	Farmacia Rosa	985 382 673	Lunes a Vi	Avda. Gaspar García	43.525911 -5.	Farmacias, 2010, Gijón				
0	http://www.gijon.es	703	es	fecha-actu	http://www.gijon.	Farmacia Pinza	985 167 544	Lunes a Vi	Casares	43.491818 -5.	Farmacias, 2010, Gijón				
1	http://www.gijon.es	702	es	fecha-actu	http://www.gijon.	Farmacia Palaci	985 386 525	rebotica@	Lunes a Viernes:9:3	Farmacia Pal	43.535841 -5.	2010, Gijón, Farmacias			
2	http://www.gijon.es	701	es	fecha-actu	http://www.gijon.	Farmacia Pagin	985 366 926	985 366 92	Lunes a Viernes:9:3	C/Avelino G	43.532453 -5.	Farmacias, 2010, Gijón			
3	http://www.gijon.es	700	es	fecha-actu	http://www.gijon.	Farmacia Molin	985 195 451	985 335 13	c08882@cofas.es	Lunes a Vier	C/Gregorio M	43.534717 -5.	Farmacias, 2010, Gijón		
4	http://www.gijon.es	699	es	fecha-actu	http://www.gijon.	Farmacia Migue	985 370 739	Lunes a Vi	C/Ezcurdia, 55	43.538236 -5.	Farmacias, 2010, Gijón				
5	http://www.gijon.es	698	es	fecha-actu	http://www.gijon.	Farmacia Merce	985 309 347	Lunes a Vi		C/Andes, 40	43.537664 -5.	Farmacias, 2010, Gijón			
6	http://www.gijon.es	697	es	fecha-actu	http://www.gijon.	Farmacia Mené	985 341 954	Lunes a Vi		C/Moros, 2	43.542002 -5.	Farmacias, 2010, Gijón			
7	http://www.gijon.es	696	es	fecha-actu	http://www.gijon.	Farmacia Mazo	985 152 191	985 386 55	Lunes a Viernes:9:3		Ctra. Obispo	43.521893 -5.	Farmacias, 2010, Gijón		

La estructura de los datos y sus tipos se representa a continuación.

Columna	Tipo	Descripción	Rango	Problemas
url	String	Representa la url de la farmacia en el directorio de Gijón.		
identificador	integer	Representa un identificador único de la farmacia.	489-1753	
locale	String	Representa el país	Siempre tiene el valor "es"	
send	String	Representa la fecha de actualización de la información.		
foto	String	Representa la foto del logo de la farmacia.		Al revisar varias la foto se repite.
nombre	String	Representa el nombre de la farmacia.		
telefono	String	Representa el número de teléfono de la farmacia.		Este dato esta en formato de texto, ya que tiene separados por un espacio grupos de 3 digitos. Además el número no esta en formato internacional.
fax	String	Representa el Fax de la farmacia.		Este dato solo esta presente en 22 de los 104 registros. Además el número no esta en formato internacional.
correo-electronico	String	Representa el correo electrónico de la farmacia.		Solo 14 registros de 104 tienen información.
web				Campo Vacio
horario	String	Representa el horario de atención de la farmacia.		
descripcion	String	Representa una descripción de lo que hace la farmacia.		Solo existe información en 30 de 104 registros.
direccion	String	Respresenta la dirección de la farmacia.		El principl problema es que la dirección no esta normalizda.
localizacion	String	Representa la geolocalización de la farmacia.		
categorias	String	Representa la categoría de la farmacia.		Casi todos los datos tienen la misma categoría.

c.- Estrategia de nombrado, donde se explique cómo se van a nombrar los recursos tanto del vocabulario a desarrollar como de los datos a generar

Para la estrategia de nombrado, es importante entender que existen 2 modelos para las URIs, estas pueden ser con # o bien con /. También es importante definir un dominio para la URIs.

- Dominio URis : http://farmaGijon.com
- Ruta para términos ontológicos: http://farmaGijon.com/ontology#
- Ruta para individuos: http://farmaGijon.com/farma/
- Patrón para términos ontológicos: http://farmaGijon.com/ontology#<term_name>
- Patrón para individuos: http://farmaGijon.com/farma/<resource_name>

d.- Desarrollo del vocabulario, indicando el proceso de implementación del vocabulario y como este soporta los datos de origen. No se exige una ontología compleja, sino un vocabulario suficiente para describir los conceptos y propiedades de los datos a transformar

Para el desarrollo del vocabulario se van a utilizar los pasos sugeridos en los videos.

- Especificar requisitos
- Extraer términos
- Elaborar conceptualización
- Buscar ontologías
- Seleccionar ontologías
- Implementar ontología
- Evaluar ontología

Especificar requisitos

Estos se dividen en requisitos funcionales (rf) y requisitos no funcionales (rfn).

rf1: Nombre de la farmacia

rf2: Página Web de la farmacia

rf3: Horario de atención de la farmacia

rf4: Dirección de la farmacia

rf5: Teléfono de la farmacia

rf6: Localización de la farmacia

rf7: Categoría de la farmacia

rnf1: La información debe tener licencia de uso http://creativecommons.org/licenses/by/3.0/es

rnf2: Debe soportar estándar

rnf3: Debe poder ser validada con herramienta Oops

rnf4: La información debe estar en español

Extraer términos

Los términos relevantes son los siguientes:

Farmacia: Establecimiento que se dedica a comercializar medicamentos

Página Web: Documento electrónico que contiene información

Teléfono: Número que representa un identificador para realizar una conexión telefónica

Dirección: Lugar donde reside la empresa o una persona

Localización: Lugar geográfico, representado por latitud y longitud

Horario: Representa un período de tiempo en el que está operativa la empresa

Conceptualización

El núcleo central es una farmacia

La farmacia tiene una página web

La farmacia tiene una dirección

La farmacia tiene un teléfono

La farmacia tiene una localización

La farmacia funciona en un horario

Buscar/Seleccionar ontologías

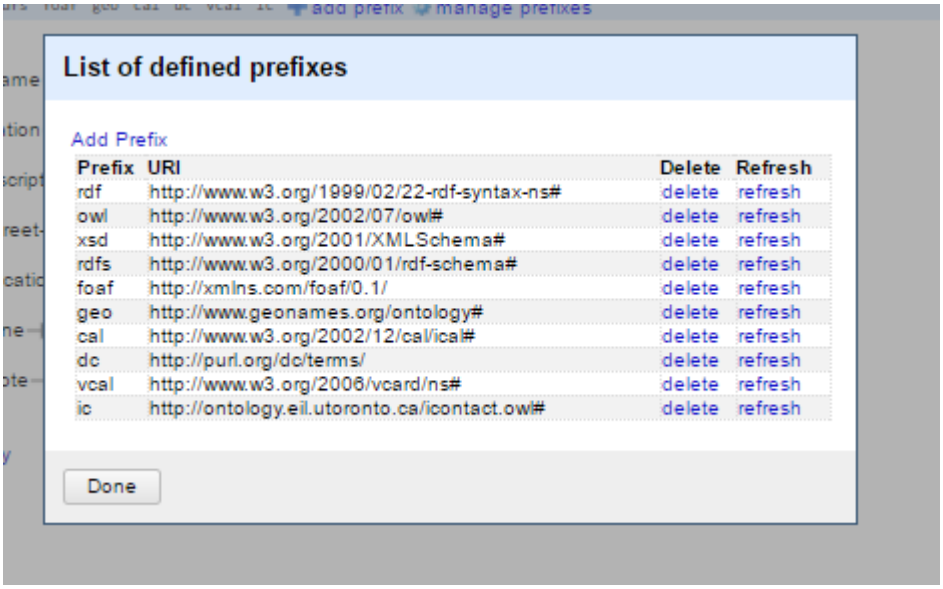
En general es bastante simple el espacio en que necesitamos buscar, una farmacia representa un lugar en el mundo.

Existe http://www.geonames.org/ontology#, que representa información geoespacial.

También existe un espacio asociado a calendar, que nos entrega vocabulario asociado: por ejemplo localización, dirección, teléfono, etc. http://www.w3.org/2002/12/cal#

Existen recursos asociados a estandarizar la meta data, estos también pueden ser utilizados para esta definición. http://purl.org/dc/terms/

En la siguiente figura se detalla los vocabularios utilizados en la transformación a RDF.



Implementar/Evaluar ontología

Dado que existen vocabularios que representan perfectamente los recursos que vamos a disponibilizar, no es necesario crear una ontología propia. En el proceso de generación del RDF se definen los vocabularios que se usan.

e.- Proceso de transformación, justificando qué herramientas se han usado para la transformación de los datos y qué pasos se han seguido para su limpieza y adecuación al resultado esperado.

El proceso seguido para la limpieza y la transformación de datos fue el siguiente:

- Rescate de información en formato CSV desde sitio web http://datos.gob.es/es/catalogo/l01330241-farmacias.
- Primer proceso de depuración utilizando Excel, en este paso se ajustaron aquellos campos que venían vacíos y desajustaban la data.
- Segundo proceso de depuración y revisión de la data utilizando OpenRefine, en este paso se eliminaron columnas que estaban vacías y alguna otra que no aportaba información, además se reviso que la data estuviera sana y consistente.
- Tercer proceso de transformación utilizando OpenRefine, esto a través de la extensión RDF.

El proyecto en OpenRefine se encuentra en el archivo "*farmacias_Gijon-xlsx.google-refine.tar*".

Un Previo en ttl se encuentra en el archivo "*farmacias_Gijon-xlsx.ttl*".

f.- Enlazado, donde se explique qué enlaces se han generado con fuentes externas y mediante qué herramientas.

Para el proceso de enlazado de datos con otras fuentes de información, fue necesario crear una columna que se llama ciudad, de manera de poder buscar otras fuentes que tengan información de esta ciudad.

Para esta nueva columna se agregó un vocabulario nuevo http://dbpedia.org/ontology/.

De esta forma podemos hacer compatible la ciudad de nuestro rdf con otros enlaces de datos.

Luego se define un reconciliation services del tipo Sparql, para esto se usa el endpoint Sparql de dbpedia español, http://es.dbpedia.org/sparql.

Se selecciona el tipo http://dbpedia.org/ontology/PopulatedPlace

Se agrega la columna PopulatePlace (para que quede referencia al tipo), la ciudad que seleccione fue Gijón, como tiene acento tuve que usar la propiedad cell.recon.best.id para acceder un valor que represente una url de la ciudad.

Luego se crea un nodo nuevo el Schema RDF con una URI que apunte al campo enlazado y se crea una propiedad que apunte a la nueva columna.

3.- Aplicación y explotación, explicando qué funcionalidades aporta la solución desarrollada y cómo ésta hace uso de los datos y enlaces generados para aportar valor al usuario final. En este punto de deben explicar las queries SPARQL o el código en Jena usado para su implementación

La aplicación entrega información acerca de las farmacias que están disponibles en Gijón, como información aparece una url, dirección, horario de atención, teléfono y un tema importante al ser utilizada desde una aplicación móvil podría mostrar la más cercana ya que tiene disponible información de localización con latitud y longitud.

Los datos no fueron publicados, pero se exportan en formato ttl. (farmacias_Gijon-xlsx.ttl)

4.-Conclusiones

El proceso de buscar información, procesarla, transformarla es bastante largo y tedioso. Al utilizar la herramienta OpenRefine este proceso se simplifica bastante. OpenRefine es una herramienta poderosa, pero requiere entrenamiento para poder tener velocidad y rendimiento en todo el proceso realizado.

La oportunidad de poder enlazar datos le otorga una potencia interesante, ya que permite muchísima colaboración y sinergia. Los estándares definidos hacen la tarea de colaboración muchísimo más fácil y permite la compatibilidad entre todas las publicaciones.

5.-Bibliografía

La bibliografía utilizada es la siguiente:

- Videos del Curso
- Material de apoyo del curso
- Documentación proporcionada por la herramienta OpenRefine
- http://lov.okfn.org