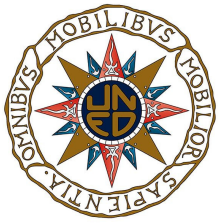
UNIVERSIDAD NACIONAL DE EDUCACION A DISTANCIA



Master de Inteligencia Artificial Avanzada: Fundamentos, Métodos y Aplicaciones.

TRABAJO FIN DE MASTER

"Recopilación guiada de datos desde Wikidata para fines educativos"

DIRECTOR:

José Luis Fernández Vindel

AUTOR:

Sergio Gutiérrez Rodríguez

Agradecimientos

Este trabajo ha sido posible gracias a la colaboración de muchas personas que con sus libros, artículos y trabajo contribuyen a la creación y mantenimiento de conocimiento y a los cuales estoy agradecido, también ha sido posible gracias a mis padres que con paciencia me han ayudado durante todo este tiempo, a mi director de proyecto José Luis Fernández Vindel con el que he mantenido numerosas reuniones y a los profesores que gracias a los materiales que ellos preparan y a su labor como docentes nos ayudan a mejorar nuestra cultura, educación y bienestar.

Resumen

En este TFM hemos realizado un trabajo de adquisición de conocimiento sobre wikidata y de creación de herramientas que utilizen recursos semánticos, en el primer capítulo realizamos una introducción sobre las nociones importantes del campo del trabajo con conocimiento. El el segundo capitulo la motivación, los objetivos en los que encontramos el conocimiento de las propiedades de wikidata y la creación de una herramienta que nos permita seleccionarlas para incorporarlas a nuestros documentos. También mostramos un capítulo que nos habla acerca de los endpoint y como podemos utilizarlos, y también un último capítulo en el que mostramos una web que permite gestionar documentos semánticos y como realizar operaciones sobre ellos con fines educativos.

Índice general

Índice de figuras

Capítulo 1

Introducción

El conocimiento es un recurso escaso cuyo refinamiento y reproducción crea bienestar.

Building Expert Systems

Para almacenar y procesar conocimiento los seres humanos han desarrollado diversas tecnologías como son los gestos, el habla, la escritura, el papel, los libros, las grabaciones de audio y vídeo, más recientemente los ordenadores, la web y los teléfonos móviles. Desde la aparición de los ordenadores los humanos no han cesado de cosechar éxitos en la resolución de problemas matemáticos y científicos que antiguamente resultaban muy difíciles de abordar.

1.1. Representación del conocimiento

Tendríamos que conocer las distintas formas de representación del conocimiento que existen para saber como almacenar conocimiento en la web y también las operaciones que podemos realizar con ese conocimiento.

1.1.1. Ordenadores y representación conocimiento

Los ordenadores son unas herramientas construidas por los humanos para almacenar y procesar información, se basan en unas operaciones lógicas y aritméticas básicas que nos permiten realizar otras operaciones más complejas con la información, nos fijamos en un ordenador elemental.

Ordenador básico

Un ordenador básico consta de la unidad aritmético lógica es un componente del ordenador en el que se realizan operaciones con la información en código binario, y se compone de componentes electrónicos digitales. La CPU reconoce instrucciones codificadas en binario que se leen de memoria e indican las operaciones que se tienen que efectuar, en memoria nos podemos encontrar tanto datos como conjuntos de instrucciones para procesar esos datos.

Cada una de las instrucciones para los humanos tienen un nombre en el lenguaje ensamblador que se corresponde con su representación en binario que es la representación que se utiliza en la electrónica digital, los registros almacenan binario, los buses de datos transportan datos en binario y la ALU realiza operaciones en binario, luego los códigos binarios se corresponden con las distintas representaciones de información que pueden ser todas representadas en binario, ya sean símbolos, letras, imágenes o sonidos.

1.1.2. Lenguaje escrito

Este es el formato más extendido en la web, y que esta orientado al caracter informativo de la web, para que los humanos puedan comunicarse leyendo y escribiendo texto escrito.

1.1.3. Lógica

La lógica es una forma de representación de conocimiento, La aceptación de los principios de la racionalidad es lo que diferencia el discurso racional de otra actividad humana. Los principios de la racionalidad distinguen un razonamiento correcto de un razonamiento no correcto. Cuando razonamos usamos pensamientos compuestos de afirmaciones primitivas sobre nuestro entorno, estas están construidas con relaciones habituales en nuestro pensamiento como, and, or, not, implies, every, some o equals. Para expresar conocimiento racional utilizamos la lógica, para la representación de conocimiento se utilizan varias lógicas distintas.Existe la logica de proposiciones que estudia que frases son correctas solo por como usan las conectivas lógicas proposicionales y la logica de predicados que añade a la lógica proposicional el uso de cuantificadores, funciones y variables.

En el caso de las logicas decidibles, los investigadores se centran en el diseño de formas de representar conocimientos que sean adecuadas tanto a nivel de expresividad como a nivel de computabilidad. Algunos de los investigadores en representación del conocimiento piensan que sería adecuado representar el conocimiento en una representación que pueda expresar objetos y las relaciones entre ellos, por ejemplo taxonomías, esta nueva forma de estructurar el conocimiento permitía a la vez representar conocimiento y realizar operaciones con el conocimiento en un tiempo adecuado. Esto dio lugar al surgimiento de los formalismos conocidos como redes semánticas y posteriormente los más avanzados sistemas basados en lógicas descriptivas.

1.1.4. Redes semánticas

Es una forma de representar la información que representa conceptos y las relaciones entre ellos. Una red semántica consiste en un grafo orientado con nodos y aristas, los nodos representan conceptos y las aristas representan relaciones entre ellos.

Las organizaciones que promueven el avance de las tecnologías web, esta promoviendo el uso de redes semánticas en internet y ya contamos con la existencia de varios estándares y herramientas para trabajar con redes semánticas en internet y en los ordenadores domésticos.

Capítulo 2

Motivación, contexto y objetivos

2.1. Motivación

Un resumen previo de este apartado sobre Motivación se puede reducir a los siguientes tres puntos:

- Existe una tecnología de estructuración de datos abiertos y de consulta de esos datos suficientemente madura
- Existe un número creciente de almacenes de este tipo de datos, sobre diversos dominios de interés. Entre estos almacenes de datos se encuentra Wikidata, que estructura la información que extrae de Wikipedia en forma de enunciados concisos
- El trabajo parte de la intuición de que el contexto digital descrito es aprovechable por los docentes para el desarrollo de actividades basadas en datos. Si tienen alguna ayuda en la descarga de datos relevantes en su asignatura, pueden implicar a grupos de estudiantes en el análisis y uso de estos datos.

2.1.1. El modelo común a todos los almacenes de datos abiertos enlazados

La página The Linking Open Data cloud diagram `http://lod-cloud.net/` actúa como indexador de los datasets públicamente accesibles por Web en forma de Datos Enlazados. En su actualización de fecha 22 de Agosto de 2017 tiene registrados 1163 datasets. Sus direcciones y descripción se pueden descargar en formato plano separado por tabuladores `http://lod-cloud.net/versions/2017-08-22/datasets_22-08-2017.tsv`. Una perspectiva más detallada puede encontrarse en LODStats `http://stats.lod2.eu/`, que ofrece una estimación de unos 150.000 millones de tripletas RDF accesibles en la nube de datos enlazados (Linked Open Data Cloud).

La introdución oficial al modelo de datos RDF se puede consultar en RDF Primer `https://www.w3.org/TR/rdf11-primer/`. Para una perspectiva más técnica, el resto de los documentos normativos sobre RDF se encuentran listados en What's new in RDF 1.1 `https://www.w3.org/TR/rdf11-new/`. Este modelo busca la precisión en la declaración de enunciados concisos sobre la relación entre datos. Se basa en una opción sintáctica y en otra que intenta evitar la ambigüedad semántica:

- El concepto de tripleta RDF como relación entre dos recursos: <sujeto><propiedad><objeto>.
- La obligación de diferenciar cada recurso de una tripleta con un identificador universal único (salvo, los <objeto>en algunos casos)

En este modelo todos los enunciados se expresan como relación entre dos recursos: <sujeto><propiedad><objeto>. Y de cada entidad, salvo los <objetos>en algunos casos, se requiere que venga nombrada por un identificador universalmente único. Como ejemplo, en

Wikidata el escritor Miguel de Cervantes Saavedra es el ítem <http://www.wikidata.org/entity/Q5682>, que puede actuar como sujeto o como objeto de un enunciado, es decir, de una tripleta o terna RDF.

2.1.2. Número y diversidad de datos disponible

2.1.3. El caso de Wikidata

Wikidata `www.wikidata.org` es uno de estos grandes almacenes de datos estructurados. Declara cerca de 43 millones de ítem en sus páginas estadísticas `www.wikidata.org/wiki/Special:Statistics`. En la terminología de este sitio, un ítem suele corresponder a una entidad o concepto que tiene página en Wikipedia. De esas páginas extrae inicialmente los datos estructurados relativos a cada ítem, que luego pueden editarse y ampliarse directamente sobre Wikidata. Este proceso de edición de datos está socialmente abierto, de forma similar a la edición de documentos en Wikipedia. Esta edición se materializa en forma de cientos o miles de tripletas RDF (<sujeto><propiedad><objeto>) por cada uno de estos ítem, sujetos de estas tripletas y potencialmente objeto en las mismas. Para empezar a familiarizarse con Wikidata se puede partir de Wikipedia. Cada página en Wikipedia tiene un item asociado en Wikidata. Supongamos, por seguir el ejemplo anterior, que estamos en la página dedicada a don Miguel de Cervantes Saavedra `https://es.wikipedia.org/wiki/Miguel_de_Cervantes`.

Hay recuadro en el margen derecho con datos estructurados. Se puede acceder al item de Wikidata correspondiente pulsando la opción 'Editar datos en Wikidata'. O bien en el menú izquierdo de esa página de Wikipedia, en la sección 'Herramientas' opción 'Elemento de Wikidata'. Así es como, por ejemplo, desde la página de Wikipedia para Miguel de Cervantes se accede a su item: Q5682. Más precisamente, Miguel de Cervantes es identificado en Wikidata con la URI `https://www.wikidata.org/entity/Q5682`. Ese identificador es Miguel de Cervantes en Wikidata (para cualquier lengua). Otra cosa son los datos que se afirman de ese identificador. Hay una página web donde se pueden modificar o ampliar estos datos. Se encuentra en `https://www.wikidata.org/wiki/Q5682`. Internamente todos estos datos acerca de Q5682 se estructuran como tripletas RDF que tienen este recurso como sujeto. Todas estas ternas se pueden descargar. Como hay más de una sintaxis que serializa RDF, se puede solicitar un volcado de esas ternas en varios formatos:

- `http://www.wikidata.org/wiki/Special:EntityData/Q5682.html`: redirige el navegador hacia el interfaz antes citado de edición web (https://www.wikidata.org/wiki/Q5682)
- `http://www.wikidata.org/wiki/Special:EntityData/Q5682.rdf` descarga las tripleta en formato rdf/xml.
- `http://www.wikidata.org/wiki/Special:EntityData/Q5682.ttl` descarga las tripletas en formato Turtle (el más conciso y legible para un agente humano)
- `http://www.wikidata.org/wiki/Special:EntityData/Q5682.nt` descarga las tripletas en formato N-Triples (todas las tripletas expandidas de forma absolutamente explícita, sin los ahorros sintácticos de Turtle); muy apropiado para su procesamiento directo por un sistema
- `http://www.wikidata.org/wiki/Special:EntityData/Q5682.json` descarga las tripletas en formato Json, apropiado para su procesamiento directamente en navegadores o en sistemas Web.

2.1.4. Sobre el potencial uso de datos descargados en el diseño de actividades formativas

Sobre Wikidata, como sobre Wikipedia, se pueden mantener reservas sobre la completud y fiabilidad de sus datos. La gran mayoría de los restantes datasets accesibles en la

Web no se editan de forma tan socialmente abierta. Usualmente son volcados de las bases relacionales internas de empresas o instituciones que actuan como indexadores sobre dominios muy diversos: desde biología, genética o medicina hasta el registro de obras (películas, libros), pasando por volcados periódicos de las oficinas estadísticas nacionales.

La posibilidad de consumo de estos datasets no está todavía popularizada. La estructuración de datos de una forma tan precisa, basada en ontologías que los jerarquizan, tiene por objetivo que haya aplicaciones inteligentes que exploten este acceso universal. Los casos de éxito que se van acumulando lo son porque se focalizan en una necesidad muy precisa de información, para una tarea concreta. y facilitan al usuario un interfaz sobre el que acotar las consultas. Y además porque incorporan conocimiento sobre los datasets que recorren.

Con esta potencialidad en mente, en este TFM se explora la línea de trabajo propuesta por el director del mismo. En su versión más genérica se plantea cómo es de factible hoy en día, técnica y metodológicamente, la recopilación de datos enlazados para su uso en el desarrollo de actividades formativas. En particular, este trabajo se centra en el análisis de Wikidata como fuente de datos. Se escogió este dataset por la diversidad de dominios de sus datos y por las facilidades de uso que, poco a poco, se van facilitando para su explotación. Facilidades que son necesarias, por otro lado, porque el desarrollo del trabajo ha ido confirmando que la conceptualización interna de Wikidata es mayor a la de muchos otros datasets, más focalizados y no coeditables.

2.1.5. Contexto y Objetivos

2.1.6. Contexto

La línea de trabajo de este TFM parte de la suposición de la viabilidad y potencial eficacia de ciertas acciones. En concreto, de las que permitirían a un docente no técnico primero recopilar datos estructurados relevantes para su asignatura, y posteriormente diseñar actividades donde grupos de estudiantes analizan y trabajan estos datos.

La usabilidad y la viabilidad dependerán de dos factores. Por un lado de que exista un sistema que haga de mediador e incorpore el conocimiento sobre la conceptualización de los datos en el dataset consultado. Por otro lado, de que este sistema disponga de un interfaz que permita construir consultas sencillas (basadas en el conocimiento del dataset) sobre un sencillo interfaz gráfico.

La potencial eficacia de esta iniciativa en el entorno docente de nuevo dependerá de dos condiciones. Por un lado, de que los datastore escogidos proporcionen datos masivos, fiables y relevantes para la asignatura en cuestión. Y por otro lado de las decisiones pedagógicas de su uso en el entorno de aprendizaje en cuestión.

Con esas consideraciones, el desarrollo de este trabajo requería cumplir algunas restricciones iniciales:

- Se precisaba escoger un dataset muy poblado de datos, que justificara el posterior guiado en la consulta. Y a ser posible, poco focalizado, variado en sus dominios de aplicación. Casi desde el principio se optó por Wikidata.

- Se requería recopilar el conocimiento sobre la diversidad y las relaciones entre todos esos datos; es decir, el modelo de datos a alto nivel del dataset.

- Con estas pistas, se pretendía recopilar algunas opciones de configuración de consultas: confirmando o descartando el interés del usuario sobre ciertas categorías de relación entre datos.

2.1.7. Objetivos

Diferenciar, por estudio del modelo de datos de Wikidata, los ejes de consulta separables que se pueden sugerir al usuario para guiar el proceso de consulta.

Definir procesos de recopilación que aprovechen este conocimiento del modelo. En particular, se pretende definir variaciones sobre un proceso genérico e iterativo de consulta donde:

- Los ejes marcados permiten configurar una primera consulta,
- Sobre los datos descargados se puede ejecutar algún proceso adicional de filtrado, reduciendo el número de ítem de interés
- El sistema permite volver a ejecutar una nueva consulta ampliatoria sobre estos ítem de interés

Desarrollar, como parte del sistema de apoyo, un entorno local donde los docentes puedan almacenar el resultado de sus diversas consultas y compartirlos entre diversos usuarios del mismo (p.ej, miembros del mismo equipo docente)

2.1.8. Metodología

En una primera aproximación parecía que el trabajo tenía tres etapas que debían completarse de forma secuencial:

- La exploración analítica del modelo de datos de Wikidata (y de su poblamiento)
- El aprovechamiento de este conocimiento para la especificación de diversos procesos guiados de consulta
- La implementación de un entorno para almacenar y coadministrar los resultados de estas consultas.

Sin embargo, la metodología finalmente utilizada ha seguido un proceso cíclico de refinamiento, de un conocimiento básico del modelo se pasó en paralelo a ejecutar algunas consultas. Y poco después se empezaba a abordar el entorno local con dos objetivos, tanto para gestionar los resultados como para disponer de un dataset local donde filtrarlo sin tener que hacerlo siempre vía refinamiento de la consulta Sparql que se enviaba a Wikidata.

La exploración analítica de Wikidata

El punto de partida de este análisis ha sido el estudio de la documentación sobre el modelo datos:

- `https://www.mediawiki.org/wiki/Wikibase/DataModel/Primer` : una introducción
- `https://www.mediawiki.org/wiki/Wikibase/DataModel`: el modelo de datos detallado.

Para asimilar este esquema de relación entre datos, se estudiaron muchos casos reales de ítem poblados con datos. Por un lado se analizaron archivos descargado de ítem diversos (especialmente en formato .ttl Turtle). Por otro lado, se analizó el resultado obtenido de diversas consultas Sparql del repositorio de ejemplos de consultas: `https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/queries/examples`. La construcción correcta de las consultas reflejan el modelo de datos en que se organiza Wikidata. A este modelo, abstracto, se ha sumado el trabajo sobre el listado que facilita Wikidata de todas las propiedades, con su tipo de dato y su frecuencia de uso: `https://www.wikidata.org/wiki/Wikidata:Database_reports/List_of_properties/all`. La creación de propiedades está moderada en Wikidata y hay poco más de cuatro mil permitidas. Y el modelo se entiende mejor y de forma más abarcable si uno se retringe inicialmente al estudio de las propiedades. Conforme se avanzaba en el conocimiento del modelo se sugerían dos líneas separadas de análisis:

- Estudio de las propiedades por su tipo de dato: sobre qué tipo de objetos eran aplicables (temporales, url, string, item, identificadores externos, ficheros multimedia ...). La determinación aquí de ejes separables de consulta se reforzaba además por la información sobre cuántas veces se usan las propiedades en ternas en Wikidata con estas restricciones.
- A esta línea de estudio hay que añadir otra complementaria, que permite clasificar a las propiedades en otros ejes: propiedades aplicables a personas, a documentos, a organizaciones, a eventos, etc.

A la clasificación combinada de estos dos criterios había que añadir el hecho de que cualquier item tiene dos conjuntos de propiedades de interés:

- Las propiedades que alojan todos los literales sobre ese ítem en los diversos idiomas de Wikipedia
- Las propiedades que enlazan ese ítem de Wikidata con todas las páginas (en los diversos idiomas) en Wikipedia que tratan sobre ese ítem.

2.2. Recopilación de todas las propiedades, con su tipo de dato y frecuencia de uso

Wikidata mantiene una página web con un volcado de su base de datos `https://www.wikidata.org/wiki/Wikidata:Database_reports/List_of_properties/all` en forma de tabla html. Esta tabla lista todas las propiedades en uso y consta de las siguientes columnas:

- Identificador de la propiedad
- Etiqueta de la propiedad (en inglés)
- Descripción de la propiedad (en inglés)
- Alias de la propiedad (en inglés)
- Tipo de dato de la propiedad
- Número de veces que la propiedad se usa en Wikidata

Por su utilidad como referencia, se ha descargado la página y se mantiene localmente esa tabla de datos. De un primer recuento sobre esta tabla se obtiene que se usan 4266 propiedades distintas en Wikidata, cada una de ellas asociada a un único tipo de dato. Ese tipo de dato es el que puede tomar el término <objeto>en el uso de una propiedad en una terna: <sujeto><propiedad><objeto>. Es decir, es el tipo de dato admisible para el rango de la relación. El objetivo de este trabajo consiste en especificar un sistema que ayude a un usuario no técnico en la recopilación de datos desde Wikidata. Uno de los criterios de búsqueda (no el único) puede serlo por el tipo de datos de la propiedad consultada. Para ese fin, esta sección analiza la tabla descargada que se mencionó al principio. Antes de eso, y para no depender del desfase de una tabla local, se enumeran las consultas Sparql que se podrían ejecutar sobre Wikidata para obtener toda esa información de partida.

2.2.1. Consultas Sparql para actualizar la tabla (propiedad, tipo de datos, frecuencia de uso)

Número de propiedades con cada tipo de datos

En la tabla citada se aprecia que hay 13 tipos de datos distintos. Por tanto, todas las propiedades se clasifican en 13 categorías distintas conforme a este criterio. Cuántas propiedades pertenecen a cada categoría es algo que puede contarse sobre la tabla descargada.

No obstante, para mantener la información actualizada en todo momento, se ha diseñado una consulta Sparql sobre Wikidata (ejecutable en `http://tinyurl.com/y9s3zq3x`):

```
# Numero de propiedades para cada uno de sus tipos de dato
SELECT (COUNT(?propiedad) as ?propNum ) ?tipoDato
WHERE {
?propiedad rdf:type wikibase:Property.
?propiedad wikibase:propertyType ?tipoDato.
}
GROUP BY ?tipoDato
ORDER BY DESC(?propNum)
```

El resultado de esta consulta es:

resultado	
propNum	**tipoDato**
2442	wikibase:ExternalId
1035	wikibase:WikibaseItem
388	wikibase:Quantity
217	wikibase:String
44	wikibase:CommonsMedia
41	wikibase:Time
40	wikibase:Url
31	wikibase:Monolingualtext
11	wikibase:WikibaseProperty
9	wikibase:Math
8	wikibase:GlobeCoordinate
4	wikibase:TabularData
2	wikibase:GeoShape

Relación de propiedades, clasificadas por su tipo de datos.

Para obtener el listado expandido de propiedades, cada una con su tipo de datos, se ha diseñado la siguiente consulta Sparql (ejecutable en `http://tinyurl.com/yaock7dy`):

```
# Todas las propiedades, con su tipo de dato
SELECT ?propiedad ?propiedadLabel ?propiedadDescription ?
tipoDato
WHERE {
?propiedad rdf:type wikibase:Property .
?propiedad wikibase:propertyType ?tipoDato .
SERVICE wikibase:label {
bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
}
ORDER BY ?tipoDato
```

El resultado de esta consulta es una tabla con las 4266 propiedades agrupadas por tipo de dato. Sus primeras filas se muestran a continuación:

resultado			
propNum	**tipoDato**		
wd:P181	mapa de distribución de un taxon	de un taxón mapa de distribución de un taxón	wikibase:CommonsMedia

continua en la próxima página

resultado			
wd:P207	imagen batimetrica	imagen del relieve del cuerpo de agua	wikibase:CommonsMedi
wd:P244	mapa de localizacion	imagen de mapa que destaca la localizacion del sujeto.	wikibase:CommonsMedi
wd:P367	símbolo astronómico	imagen del símbolo que identifica a un planeta o asteroride del sistema solar	wikibase:CommonsMedi

Número de veces que se usa cada propiedad en Wikidata.

Las dos consultas previas permiten a nuestro sistema recabar casi toda la información disponible en el volcado que se mencionó al principio. Toda, salvo el número de veces que se usa cada propiedad en Wikidata, porque una consulta global así colapsa el servicio. Se puede programar por separado para cada propiedad, tal y como se ve en la siguiente consulta (ejecutable en `http://tinyurl.com/y6wgdnys`):

```
# Numero de usos de una propiedad en Wikidata
# Aqui, p.ej., la P20 ('sitio de fallecimiento')
SELECT (COUNT (?item) AS ?itemNum)
WHERE {
?item wdt:P20 ?sitio .
}
```

En cuanto al número de usos en Wikidata de todas estas propiedades, se obtiene una suma total de 364.008.654 ocurrencias en el dataset, con una distribución desigual. De hecho, las propiedades que superan el millón de usos son 46 y la suma total de las ocurrencias de estas 46 propiedades en el dataset llega a 297.720.379

resultado			
ID	**Label**	**Data type**	**Count**
P31	instance of	wikibase-item	38956316
P248	stated in	wikibase-item	20238436
P813	retrieved	time	8268667
P577	publication date	time	15085813
P143	imported from	wikibase-item	13704825
P1476	title	monolingualtext	13689164
P1433	published in	wikibase-item	13155665
P1545	series ordinal	string	12834813
P304	page(s)	string	12771148
P2093	author name	string	12763193
P478	volume	string	12517045
P698	PubMed ID	external-id	12511616
P433	issue	string	11203126
P356	DOI	external-id	9834092
P17	country	wikibase-item	8341773
P854	reference URL	url	5760922
P625	coordinate location located in the administrative territorial	globe-coordinate	5363415

continua en la próxima página

resultado			
P131	entity	wikibase-item	4970606
P932	PMCID	external-id	3885632
P21	sex or gender	wikibase-item	3609995
P569	date of birth	time	2980459
P1566	GeoNames ID	external-id	2960116
P106	occupation	wikibase-item	2608192
P735	given name	wikibase-item	2552722
P225	taxon name	string	2363349
P105	taxon rank	wikibase-item	2363312
P171	parent taxon	wikibase-item	2363207
P27	country of citizenship	wikibase-item	2191806
P18	image	commonsMedia	1940673
P407	language of work or name	wikibase-item	1816062
P373	Commons category	string	1790167
	Global Biodiversity Information Facility		
P846	ID	external-id	1735399
P19	place of birth	wikibase-item	1537679
P2860	cites	wikibase-item	1498796
P2044	elevation above sea level	quantity	1537679
P570	date of death	time	1498796
P1435	heritage designation	wikibase-item	1431352
P421	located in time zone	wikibase-item	1400686
P830	Encyclopedia of Life ID	external-id	1375317
P279	subclass of	wikibase-item	1347793
P50	author	wikibase-item	1336990
P703	found in taxon	wikibase-item	1264532
P646	Freebase ID	external-id	1258185
P351	Entrez Gene ID	external-id	1228564
P921	main subject	wikibase-item	1168861
P214	VIAF ID	external-id	1160806
P352	UniProt protein ID	external-id	985115
P361	part of	wikibase-item	906388

2.2.2. Primeros resultados del análisis de estas tablas

Se listan 4266 propiedades distintas, agrupadas en 13 categorías (conforme a su tipo de datos). La categoría mayoritaria (ExternalID, con 2442 propiedades) es la de las propiedades que sirven para enlazar un item a su identificador en otra base externa de datos enlazados; por ejemplo <Cervantes><tiene por identificador en la Biblioteca Nacional de España><XXXX>. La siguiente categoría (WikibaseItem, con 1035 propiedades) es más heterogénea: incluye todas las propiedades que relacionan un recurso con identificador en Wikidata con otro; por ejemplo <Cervantes><tiene por lugar de nacimiento><Alcalá de Henares>. De las otras 11 categorías interesarán especialmente las que fijen ejes de consulta diferenciados; por ejemplo, las que se usan para complementar datos temporales (Time) o geográficos (GlobeCoordinate) asociados a un ítem. O bien archivos multimedia internos asociados al mismo (CommonsMedia), o enlaces genéricos a direcciones web relativas (Url). El guiado en las consultas, considerando estos ejes, sugiere que el sistema permita al usuario (1) decidir si desea o no información sobre identificadores externos o sobre archivos multimedia en CommonsMedia (2) navegar por las propiedades con tipo de dato Wikiba-

seItem y (3) utilizar la información temporal o geográfica para filtrar el conjunto de datos recibidos o para disparar una nueva consulta de ampliación de los mismos. Adicionalmente, si el sistema mantiene la información sobre frecuencia de uso de las propiedades puede sugerir en la ampliación de consultas aquellas propiedades más frecuentes (aplicables en ese contexto determinado).

Propiedades con tipo ExternalID

Las propiedades que tienen por tipo de dato Wikibase:ExternalID interesan a este trabajo porque pueden permitir al sistema ofrecer al usuario enlaces hacia fuentes externas estructuradas. Enlaces donde ampliar información sobre un ítem de nuevo mediante consultas Sparql en esos datasets. Consultas que pueden ser incluso federadas: es decir, que pueden mezclar información de más de un dataset. Estos enlaces se producen porque alguno de los editores de datos declara que un ítem de Wikidata es semánticamente el mismo concepto que el que representa otro ítem en una base de datos enlazados externa. Por ejemplo, el identificador `http://www.wikidata.org/entity/Q5682` en Wikidata referencia al mismo escritor (Miguel de Cervantes) que el identificador `http://datos.bne.es/persona/XX1718747.html` en la Biblioteca Nacional de España. La terna que declara esto en el dataset de Wikidata es (abreviada) <Q5682><tiene por BNE ID><XX1718747>. Así, la propiedad <tiene por BNE ID>sólo se usa para enlazar a ese dataset. Es decir, hay tantas propiedades del tipo ExternalID como dataset externos se estén referenciando desde Wikidata. Para cada dataset nuevo que se referencie hay que añadir una propiedad para ello. Ya se mencionó que hay 2442 propiedades de este estilo. Otro tema es cuántas referencias distintas se producen desde Wikidata a un mismo dataset externo. La siguiente tabla presenta las 15 propiedades de esta categoría más utilizadas. Suman 39.954.985 apariciones del total de 61.745.255 de las presentadas por propiedades de esta categoría.

propiedades		
P698	PubMed ID	12511616
P356	DOI	9834092
P932	PMCID	3885632
P1566	GeoNames ID	2960116
P846	Global Biodiversity Information Facility ID	1816062
P830	Encyclopedia of Life ID	1375317
P646	Freebase ID	1258185
P351	Entrez Gene ID	1228564
P214	VIAF ID	1160806
P352	UniProt protein ID	985115
P442	China administrative division code	742519
P227	GND ID	592821
P244	Library of Congress authority ID	543599
P345	IMDb ID	531338
P213	ISNI	529203

Una aproximación a la distribución de enlaces externos se puede apreciar en la siguiente lista. Resume el número de datastore externos clasificados por el total de enlaces que referencian desde Wikidata (via la propiedad correspondiente) a cada uno:

- >1.000.000, 9
- Entre 100.000 y 1.000.000, 52
- Entre 20.000 y 100.000, 166

- Entre 5.000 y 20,000, 239
- Entre 1.000 y 5.000, 495
- Entre 300 y 1.000, 416
- Entre 100 y 300, 318
- Entre 0 y 100, 744

Uso de este tipo de propiedades hacia datasets externos

En el sistema de consulta que estamos perfilando, este es uno de los ejes de búsqueda que deben activarse o no en el interfaz. En cualquier consulta que delimite un conjunto de ítem se puede adicionalmente solicitar sus enlaces a una base de datos externa (o más generalmente, a todas las que referencien esos ítem). Como ejemplo se muestra la siguiente consulta Sparql (ejecutable en `http://tinyurl.com/y9l8om7s`) que además de acotar todas las universidades públicas de España (registradas en Wikidata) solicita el identificador de esas universidades en Freebase (tal y como consta registrado en Wikidata):

```
# Universidades publicas (Q875538) en España (Q29)
# que 'tengan por identificado en Freebase" (P646) ...
SELECT ?univ ?univLabel ?freeBaseID
WHERE {
?univ wdt:P31 wd:Q875538 .
?univ wdt:P17 wd:Q29 .
OPTIONAL { ?univ wdt:P646 ?freeBaseID .}
SERVICE wikibase:label {
bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
}
```

En una línea similar se pueden obtener todos los datasets externos a los que referencia un ítem o un conjunto de ítem dados. Por ejemplo, la siguiente consulta Sparql (ejecutable en `http://tinyurl.com/y8hruuzl`) lista todos las referencias a datasets externos del ítem UNED en Wikidata:

```
# Del item UNED (Q421739), relación de enlaces a sus
identificadores en datasets externos
# tal y como se mantienen registrados en Wikidata
SELECT ?propiedadLabel ?propiedadDescription ?valor WHERE {
wd:Q421739 ?propEnUso ?valor.
?propiedad wikibase:directClaim ?propEnUso.
?propiedad wikibase:propertyType <http://wikiba.se/
ontology#ExternalId> .
SERVICE wikibase:label {
bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
}
```

Las propiedades del tipo CommonsMedia

Al igual que los identificadores externos, este tipo de propiedades caracteriza un tipo de información que se desea o no incluir en la consulta. Puede que se esté buscando específicamente material gráfico o multimedia sobre un conjunto de ítem. O bien que toda esta categoría de datos sea irrelevante para el tipo de consulta que se está ejecutando. Por tanto, estos dos ejes (identificadores externos y multimedia asociada) se pueden singularizar en el interfaz del sistema para permitir o no que el filtrado o la ampliación de datos los incluyan. Las propiedades que enlazan a recursos gráficos en CommonsMedia son pocas. Se listan a continuación tal y como aparecen en el volcado de propiedades antes citado:

propiedades					
ID	**Label**	**Description**	**Aliases**	**Data type**	**Count**
P18	image	image of relevant illustration of the subject; if available, use more specific properties (sample: coat of arms image, locator map, flag image, signature image, logo image, collage image); only images which exist on Wikimedia Commons are acceptable	portrait, illustration, picture, drawing, photo, diagram, img	commonsMedia	2.246.535
P242	locator map image	geographic map image which highlights the location of the subject	map, locator map	commons media	138093
P94	coat of arms image	image of the item's coat of arms	CoA image	commonsMedia	92.702
P154	logo image	graphic mark or emblem commonly used by commercial enterprises, organizations and products	insignia, sign	commonsMedia	42.965
P41	flag image	image of the item's flag	image flag, image of flag	commonsMedia	30.628
P1442	image of grave	picture of a person or animal's grave, gravestone or tomb	grave image, grave photo, grave picture, headstone image, headstone photo, headstone picture, Image of Shrine, Image of Tomb	commonsMedia	24.052
P1943	location map	location map of place		commonsMedia	21.427

continua en la próxima página

propiedades					
P948	page banner	a lead image about the topic, mainly used by Wikivoyages	Wikivoyage banner	commonsMedia	16.691
P15	route map	image of route map at Wikimedia Commons	street map, road atlas, highway map, schema, road map, railway map, railroad map, map of route, metro map, subway map, underground map	commonsMedia	11.589
P109	signature	image of a person's signature	autograph	commonsMedia	11.326
P14	graphic symbol of throughfare	graphic representing the thoroughfare	motorway sign, shield, highway shield, road marker, route marker, highway marker, trail blaze	commonsMedia	9.856
P181	taxon range map image	range map of a taxon	spread map, distribution map (taxa), range map	commonsMedia	8.918
P692	gene atlas image	image showing the GeneAtlasexpression pattern		commonsMedia	8.742
P1801	commemorative plaque image	plaque image, memorial plaque image, inscription commonsMe image		commonsMedia	8.377
P117	chemical structure	image of a representation of the structure for a chemical compound	crystallographic structure	commonsMedia	5.629

continua en la próxima página

propiedades					
P996	scanned file on wikimedia commons	file on Wikimedia Commons related to the content of the source/book/-report	scan file (Commons), document file on Wikimedia Commons, full text, sheet music, full movie, book image, book file, scan image	commonsMedia	4.859
P443	pronunciation audio	audio file with pronunciation		commonsMedia	3.499
P158	seal image	image of subject's seal (emblem)	emblem	commonsMedia	2.984
P51	audio	relevant sound	song, recording, audio file, audio recording, sound recording	commonsMedia	2.203
P2425	service ribbon image	an image depicting the ribbon associated with a medal, order, etc.	medal ribbon image, ribbon image	commonsMedia	1.996
P10	video	relevant video	media	commonsMedia	1.598
P1766	place name sign	image of road sign with place name on it city limit sign		commonsMedia	1.149
P1621	detail map	map containing details about the entire location		commonsMedia	1.079
P3451	nighttime view	image of the subject at night, or at least night image, in twilight	image at night, night image, night time view	commonsMedia	758
P990	audio recording of the subjects spoken voice	audio file representing the speaking voice of a person; or of an animated cartoon or other fictitious character	voice recording, recording of the subject's spoken voice, spoken voice, speaking voice	commonsMedia	707

continua en la próxima página

propiedades					
P989	spoken text audio	Wikipedia article, including audio descriptions		commonsMedia	613
P1846	distribution map	distribution of item on a mapped area (for range map of taxa, use (P181).)		commonsMedia	493
P2910	icon	pictogram suitable to represent the item. For logos of an organization, use "logo pictogram, image"(P154)	pictogram, symbol	commonsMedia	409
P1944	relief location map	relief location map of place		commonsMedia	344
P3383	film poster	poster used to promote and advertise this film (if file is available on Commons). Use P154 for logos, P18 for movie stills and related images. Items about film posters can be linked with the movie poster qualifier "subject of"(P805).	movie poster, poster of film	commonsMedia	326

continua en la próxima página

propiedades					
P1543	monogram	image of a person's monogram		commonsMedia	188
P207	bathymetry image	image showing bathymetric chart, bathymetric map	bathymetric chart, bathymetric map, file bathymetry	commonsMedia	182
P4291	panorama view	panorama view of the object	panoramic view	commonsMedia	157
P3311	plan image	image representing the plan of a building or place	plan view, airport diagram, architectural plan image	commonsMedia	125
P2716	collage image	image file that assembles two or more other images of item		commonsMedia	93
P2919	label in sign language	media file showing label of this item in sign language. Use "language of work or name"(P407) as qualifier to indicate which language.		commonsMedia	77
P2343	playing range image	image showing the playing range of the range image instrument		commonsMedia	43
P2713	sectional view	image file that shows a sectional view of the item		commonsMedia	41
P367	astronomic symbol image	image of the symbol that identify a planet or an asteroid of the solar system official shield image; see also P41, P158and P94		commonsMedia	33

continua en la próxima página

propiedades					
P4004	shield image	official shield image; see also P41, P158 shield image and P94	shield	commonsMedia	15
P3030	sheet music	media file containing the musical score (sheet music) for this item	musical score	commonsMedia	13
P4640	photosphere image	image with the field of view 360x180 image degrees		commonsMedia	13
P491	orbit diagram sandbox	image with the diagram of an orbit diagram astronomic body		commonsMedia	10
P368	commons media file	Sandbox property for value of type Çommons Media File"		commonsMedia	4

También se pueden obtener, con los literales en español, a través de la siguiente consulta Sparql (ejecutable en `http://tinyurl.com/ybfhs943`):

```
# Todas las propiedades del tipo CommonsMedia
SELECT ?propiedad ?propiedadLabel ?propiedadDescription
WHERE {
?propiedad wikibase:propertyType wikibase:CommonsMedia .
SERVICE wikibase:label {
bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
}
```

Para solicitar de un item (o conjunto de ellos) todos los recursos en CommonsMedia se puede ejecutar una consulta similar a ésta, donde se pregunta por todos estos recursos asociados a la UNED (ejecutable en `http://tinyurl.com/ycd37587`):

```
# Todos los recursos multimedia asociados a la UNED en CommonsMedia
SELECT ?propiedadID ?propiedadIDLabel ?recurso ?recursoLabel
WHERE {
wd:Q421739 ?propiedadUso ?recurso .
?propiedadID wikibase:directClaim ?propiedadUso.
?propiedadID wikibase:propertyType wikibase:CommonsMedia .
SERVICE wikibase:label {
bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
}
```

Y que da como resultado dos archivos:

Resultado			
propiedad ID	**propiedad IDLabel**	**recurso**	**recursoLabel**
continua en la próxima página			

Resultado			
wd:P154	logotipo	commons:LogoUNED.jpg	http:// commons.wikimedia.org/wiki/Special:FilePath/LogoUNED.jpg
wd:P94	imagen del escudo de armas	commons:EscudoUN ED.jpg	http:// commons.wikimedia.org/wiki/Special:FilePath/EscudoUNED.jpg

Las propiedades del tipo WikibaseItem

Estas propiedades relacionan dos ítem en Wikidata: dos recursos que tiene URI asignado y que muy posiblemente tengan ambos página en Wikipedia. En este momento son 1031 propiedades que relacionan ítem muy dispares. No es factible aquí desconectar o no este eje de consulta en su conjunto, como sí que se podía sugerir con los identificadores externos o los recursos en CommonsMedia. Del volcado que facilita Wikidata se puede calcular que este tipo de propiedades se usan en 153.932.782 tripletas. Se listan las propiedades más frecuentes (por encima del medio millón de usos).

Propiedades			
Label	**Description**	**Data type**	**Count**
instance of	that class of which this subject is a particular example and member. (Subject typically an individual member with Proper Name label.) Different from P279 (subclass of)	wikibase-item	38.956.316
stated in	to be used in the references field, to the information source in which a claim is made was made; for qualifiers use P805	wikibase-item	20.238.436
imported from	source of this claim's value (use only in References section)	wikibase-item	13.704.825
published in	larger work that a given work was	wikibase-item	13.155.665
country	sovereign state of this item	wikibase-item	8.341.773
located in the administrative territorial entity	the item is located on the territory of the following administrative entity. Use P276 (location) for specifying the location of non-administrative places and for items about events	wikibase-item	4.970.606
sex or gender	sexual identity of subject: male (Q6581097), female (Q6581072), intersex (Q1097630), transgender female (Q1052281), transgender male (Q2449503). Animals: male animal (Q44148), female animal (Q43445). Groups of same gender use "subclass of"(P279)	wikibase-item	3.609.995

continua en la próxima página

Propiedades			
occupation	occupation of a person; see also "field of work"(Property:P101), "position held"(Property:P39) first name or another given name of this person;	wikibase-item	2.608.192
given name	first name or another given name of this person; values used with the property shouldn't link disambiguations nor family names	wikibase-item	2.552.722
taxon rank	level in a taxonomic hierarchy	wikibase-item	2.363.312
parent taxon	closest parent taxon of the taxon in question	wikibase-item	2.363.207
country of citizenship	the object is a country that recognizes the subject as its citizen	wikibase-item	2.332.557
language of work or name	language associated with this work or name (for persons use P103 and P1412)	wikibase-item	2.191.806
place of birth	most specific known (e.g. city instead of country, or hospital instead of city) birth location of a person, animal or fictional character	wikibase-item	1.790.167
cites	citation from one creative work to another	wikibase-item	1.735.399
heritage designation	heritage designation of a historical site	wikibase-item	1.431.352
designation located in time zone	time zone for this item	wikibase-item	1.400.686
subclass of	all instances of these items are instances of those items; this item is a class (subset) of that item. Not to be confused with P31 (instance of)	wikibase-item	1.347.793
author	main creator(s) of a written work (use on works, not author humans	wikibase-item	1.336.990
found in taxon	the taxon in which the item can be found	wikibase-item	1.264.532
main subject	primary topic of a work	wikibase-item	1.168.861
part of	object of which the subject is a part. Inverse property of "has part"(P527). See also "has parts of the class"(P2670).	wikibase-item	906.388
sport	sport in which the entity participates or belongs to	wikibase-item	835.541
location	location of the item, physical object or event is within In case of an administrative entity use P131. In case of a distinct terrain feature use P706.	wikibase-item	807.951

continua en la próxima página

Propiedades			
country of origin	country of origin of the creative work or subject item	wikibase-item	761.204
place of death	most specific known (e.g. city instead of country, or hospital instead of city) death location of a person, animal or fictional character	wikibase-item	685.200
languages spoken writen or signed	language(s) that a person speaks or writes, including spoken,the native language(s)	wikibase-item	586.253
genre	a creative work's genre or an artist's field of work genre (P101). Use main subject (P921) to relate creative works to their topic	wikibase-item	567.847
strand orientation	orientation of gene on double stranded DNA molecule	wikibase-item	538.389
determination method	How a value is determined, or the standard method by which it is declared	wikibase-item	534.921
family name	surname or last name of a person	wikibase-item	514.073
encoded by	the gene that encodes some gene product	wikibase-item	513.678
encodes	the product of a gene (protein or RNA)	wikibase-item	511.458

Si se pretende que el sistema consulte todos los ítem relacionados con uno dado, se puede conseguir de forma similar a la siguiente consulta Sparql (ejecutable en `http://tinyurl.com/yakz5cca`) donde se obtienen los ítem relacionados con la UNED:

```
# Todos los item (con página en Wikipedia) asociados a la UNED
SELECT ?propiedadID ?propiedadIDLabel ?recurso ?recursoLabel
WHERE {
wd:Q421739 ?propiedadUso ?recurso .
?propiedadID wikibase:directClaim ?propiedadUso.
?propiedadID wikibase:propertyType wikibase:WikibaseItem .
SERVICE wikibase:label {
bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
}
```

Y que tiene por resultado.

propiedades			
propiedad ID	**propiedad IDLabel**	**recurso**	**recursoLabel**
wd:P17	pais	wd:Q29	España
wd:P31	instancia de	wd:Q875538	universidad pública
wd:P101	campo de trabajo	wd:Q159595	educación a distancia
wd:P159	ubicación de la sede	wd:Q2807	Madrid
wd:P910	categoría principal del tema	wd:Q9035711	Categoría: Universidad nacional de educación a distancia.

2.3. Propiedades, por su frecuencia de uso

Número de veces que se usa cada propiedad en Wikidata. Las dos consultas previas permiten a nuestro sistema recabar casi toda la información disponible en el volcado que se mencionó al principio. Toda, salvo el número de veces que se usa cada propiedad en Wikidata, porque una consulta global así colapsa el servicio. Se puede programar por separado para cada propiedad, tal y como se ve en la siguiente consulta (ejecutable en `http://tinyurl.com/y6wgdnys`):

```
# Numero de usos de una propiedad en Wikidata
# Aqui, p.ej., la P20 ('sitio de fallecimiento')
SELECT (COUNT (?item) AS ?itemNum)
WHERE {
?item wdt:P20 ?sitio .
}
```

En cuanto al número de usos en Wikidata de todas estas propiedades, se obtiene una suma total de 364.008.654 ocurrencias en el dataset, con una distribución desigual. De hecho, las propiedades que superan el millón de usos son 46 y la suma total de las ocurrencias de estas 46 propiedades en el dataset llega a 297.720.379.

2.4. Propiedades, por su taxonomía

2.4.1. Jerarquización transitiva de las propiedades

El esquema RDF del W3 reserva un identificador para declarar que una propiedad es subpropiedad de otra: <propiedad1><es subpropiedad de><propiedad2>. Es decir, que cualesquiera dos recursos relacionados por la <propiedad1>también están relacionados por la <propiedad2>. En Wikidata, la propiedad 'subproperty of' (P1647) se define como equivalente a la del esquema RDF citada. El uso de esta propiedad P1647 en Wikidata produce una jerarquización (transitiva) de sus propiedades: una propiedad A es subpropiedad de otra B, que a su vez lo es de otra C. Y por lo tanto, la propiedad A lo es también de C aunque no conste explícitamente así en el dataset. Mediante el uso de una expresión compleja (de un property path) en SPARQL se puede preguntar de golpe por todas las propiedades de las que una determinada es transitivamente subpropiedad. Basta utilizar la expresión regular P1647+ . Un sistema que conozca esta jerarquización de las propiedades en Wikidata puede utilizarla para ayudar al usuario a filtrar resultados o a ampliarlos siguiendo ciertos ejes determinados.

Clasificación de la aplicabilidad de las propiedades

La jerarquización anterior, como subpropiedades, agrupa cadenas de propiedades. En cada caso, la que parte de una propiedad sin subpropiedades y recorre hacia arriba todas las propiedades que la engloban hasta llegar a la última de la cadena, sin propiedad padre. Wikidata tiene otro mecanismo de agrupación de propiedades, para su clasificación, atendiendo al uso de las mismas. Para ello se proponen ítem como, p. ej. Q18608871 ('propiedad de Wikidata para elementos sobre personas') y se declaran como instancias suyas todas la propiedades que relacionan personas con otros recursos: <persona><propiedad><información sobre personas>. Aunque la jerarquización antes mencionada y esta clasificación (por instanciación) sean distintas, es obvio que si es instancia de Q18608871 la propiedad superior (raíz) de una de las cadenas citadas también lo serán todas sus descendientes: todas se aplican a personas y las relacionan con un determinada categoría de recurso. Ahora bien, pueden ser instancias de este ítem Q18608871 otras cadenas de propiedades distintas: las que también se apliquen a personas pero las relacionen con otra categoría de recursos. Si nos olvidamos de la jerarquización como subpropiedades,

la mera clasificación de propiedades por su aplicabilidad es también un conocimento que cualquier sistema de apoyo a la consulta en Wikidata puede explotar.

Una consulta exploratoria que mezcla las dos conceptualizaciones citadas

Para continuar con la exploración analítica de este dataset se ha diseñado una consulta Sparql (ejecutable en `http://tinyurl.com/y8dnf2su`) que conjuga los dos criterios mencionados. Para cada propiedad analizada (todas en Wikidata) expande todas las propiedades de las que es subpropiedad (de forma explícita y directa o transitivamente). Y cada una de estas filas las replica tantas veces como clases distintas tienen a la propiedad analizada como instancia suya. Es decir, si se presentaran estos resultados como árboles expandibles se tendrían tantos árboles como propiedades en Wikidata. Y en cada uno de ellos la primera expansión mostraría tantas ramas como propiedadas que incluyan a la raíz. Y en cada uno de estos nuevos nodos, la expansión muestra todos los ítem de los que es instancia la propiedad raíz del árbol.

```
# De cada propiedad se solicita la propiedad,
# si la tuviera, de la que es subpropiedad
# (P1647);
# de hecho, se solicita el cierre transitivo (P1647+):
# propiedades de las que es subsubsubsub...propiedad.
# En paralelo, tambien de cada propiedad
#  se solicita de que item es instancia (P31),
# o mas concretamente instancia de alguna subsub...
#  clase del mismo (P31/P279*);
# esto es un artificio en Wikidata para clasificarlas.
# El numero de filas resulta del producto
# cartesiano de ambos resultados,
# que se podria dividir por cuatro porque toda
# propiedad es instancia de
# 'Wikibase item', 'objeto', 'objeto abstracto' y 'entidad'.

SELECT
?propiedad ?propiedadLabel ?subPropiedadDe ?subPropiedadDeLabel ?
instanciaDe ?instanciaDeLabel
WHERE {
?propiedad rdf:type wikibase:Property.
OPTIONAL {?propiedad wdt:P1647+ ?subPropiedadDe .}
OPTIONAL {?propiedad wdt:P31/wdt:P279* ?instanciaDe.}
SERVICE wikibase:label { bd:serviceParam wikibase:language
"[AUTO_LANGUAGE],en". }
}
ORDER BY ASC(?propiedad) ASC(?subPropiedadDe) ASC(?instanciaDe)
```

Capítulo 3

Explorando un endpoint

Para explorar endpoints sería adecuado seguir un procedimiento habitual para conocer la estructura del contenido semántico del endpoint y también poder encontrar la información que buscamos de una manera sencilla.

Normalmente en los endpoint sparql hay un conjunto de relaciones taxonómicas habituales rdf:subclassof rdf:type, owl:class, owl:subclass.

No obstante la forma más adecuada o más rápida de empezar la consulta es utilizando la función regex para buscar los conceptos que tienen relación con lo que estamos pensando.

```
select * where {?s ?p ?o .
        FILTER(regex(?o , "cat", "i"))
        } limit 100
```

Para explorar los nodos y arista a la derecha del nodo se explora de la siguiente forma.

```
select ?edge ?rightnode where
               {
                <nodo> ?edge ?rightnode .
               }
```

Para explorar los nodos y arista a la izquierda del nodo se explora de la siguiente forma.

```
select ?leftnode ?edge where
               {
               ?leftnode  ?edge  <nodo> .
               }
```

3.1. Almacenamiento de conocimiento en wikidata

Haremos un pequeño estudio de como wikidata representa distintos conceptos que podemos encontrarnos dentro de los conocimientos habitualmente publicados en la web.

Vamos a centrarnos en areas de conocimiento.

3.1.1. Fauna

Nos centraremos en el ratón o también conocido como Mus musculus Buscando las relaciones posibles de ratón común nos encontramos 2 relaciones interesantes.

```
select distinct ?edge ?edgeLabel where
{
?leftnode  ?edge  wd:Q83310 .
SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
}
```

```
ps:P279          http://www.wikidata.org/prop/statement/P279
pq:P703          http://www.wikidata.org/prop/qualifier/P703
ps:P703          http://www.wikidata.org/prop/statement/P703
ps:P31   http://www.wikidata.org/prop/statement/P31
ps:P171          http://www.wikidata.org/prop/statement/P171
ps:P921          http://www.wikidata.org/prop/statement/P921
ps:P1531         http://www.wikidata.org/prop/statement/P1531
pq:P2352         http://www.wikidata.org/prop/qualifier/P2352
...
```

Son las propiedades que aparecen en el resultado de esas dos las mas relevantes para el conocimiento de las especies son las siguientes.

```
ps:P703 found in taxon.
ps:P171 parent taxon.
```

Buscando las nodos de los cuales es padre taxonómico el Mus musculus.

```
select distinct ?leftnode ?leftnodeLabel  where
{
?leftnode  ps:P171  wd:Q83310 .
SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
}
```

```
 wds:Q1957638-00F55B31-9E89-4C76-B2C1-AFEFC2CDD1A7
wds:Q40539762-C90C840B-741F-4A12-BA51-389063CD6192
wds:Q2683493-AC3FAFCB-9D57-4117-9707-3322B03B3862
wds:Q3061141-2765A6F7-ED80-4925-A7B8-C999C8128154
wds:Q20904855-401E656F-A1B6-4CB3-8F44-E270535E9F45
wds:Q3867458-1DD09BB9-FFCF-4381-8EF7-66A60F722F81
wds:Q20904857-B79926FC-A1B1-43E2-95DF-9DAC0806FB06
...
```

Lo cual indica que wikidata si tiene una taxonomía de especies bastante poblada lo observaremos para otra especie.

Ahora observaremos para la ardilla, Sciuridae.

```
select distinct ?leftnode ?leftnodeLabel where
{
?leftnode  ps:P171  wd:Q9482.
SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
}
```

```
wds:Q19866036-aeb87925-45bf-2f76-c201-af7c84a7da09
wds:Q30046827-bc21647e-4266-f016-bcac-9827692a2db2
wds:Q27486015-9b77438d-41d8-1485-5fb5-1991695e4409
wds:Q33140837-DFDC9AD8-16FC-4F86-A455-97812BC795EF
```

```
wds:Q33176195-56856F65-8203-4B68-A77C-A5B29161807F
wds:Q33176935-642BAD46-77CF-49ED-B8BA-FDBBEF235CA6
wds:Q33176961-3D639BA5-63A1-4ECA-9260-7B9B6866ADE4
wds:Q20972224-824c6adf-4e07-e71c-e412-702a86f5ef72
...
```

también observamos algunos resultados.

La parte de taxonomía de especies esta cubierta.

3.1.2. Oceanos

Una forma sencilla de buscar los identificadores de los conceptos en wikidata es utilizando el buscador de wikidata ya que nos contesta con los resultados más habituales.

Esta vez vamos a observar los cuerpos de agua empezaremos buscando el mediterraneo.

La siguiente consulta nos deja ver las propiedades.

```
select distinct ?edge ?edgeLabel where
{
?leftnode ?edge wd:Q4918 .
SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE
}
```

```
ps:P276          http://www.wikidata.org/prop/statement/P276
ps:P361          http://www.wikidata.org/prop/statement/P361
ps:P403          http://www.wikidata.org/prop/statement/P403
ps:P706          http://www.wikidata.org/prop/statement/P706
ps:P20  http://www.wikidata.org/prop/statement/P20
ps:P119          http://www.wikidata.org/prop/statement/P119
ps:P183          http://www.wikidata.org/prop/statement/P183
ps:P180          http://www.wikidata.org/prop/statement/P180
...
```

de todas estas propiedades nos interesan las siguientes:

```
ps:P706  located on terrain feature
ps:P527  has part
ps:P361  part of
ps:P276  location
ps:P206  located next to body of water
```

Estas propiedades describen las características geográficas del mar mediterraneo y es una forma de describirlo bastante completa lo que vuelve a ser una exitosa realización de wikidata que tiene bastante completa su información geográfica.

3.2. Consultas en Wikidata

Podemos ver la imagen del endpoint de wikidata en el que se pueden hacer consultas, también se pueden hacer consultas a un servicio web que wikidata a habilitado para tal fin.

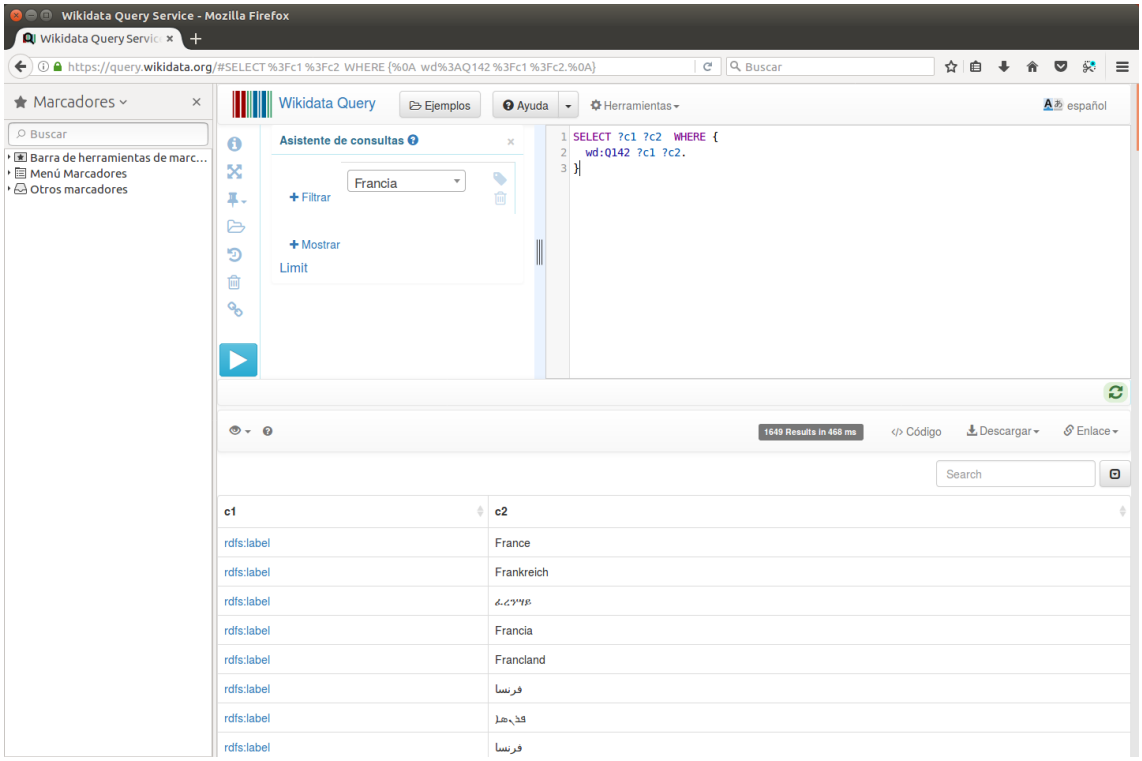


Figura 3.1: Explorar fichero web

Como ejemplo hemos realizado algunas consultas que muestran como se puede utilizar sparql para consultar wikidata.

- Contar los paises que tienen frontera con Francia:

```
select (count(distinct ?c2) as ?count)
where {
                    wd:Q142   wdt:P47 ?c2
                    }
```

- Contar los territorios que están situados cerca del cuerpo de agua del atlántico

```
select   (count(distinct ?c1) as ?count)
WHERE {
                    ?c1  wdt:P206  wd:Q97.
                    }
```

- Mostrar los territorios que están situados cerca del cuerpo de agua del atlántico y del pacífico.

```
select   ?c1
WHERE {
                    ?c1  wdt:P206  wd:Q97;
                wdt:P206  wd:Q98.
        }
```

- Mostrar los territorios que estén situados cerca del cuerpo de agua oceano atlántico pero no del pacífico.

```
select   ?c1
WHERE {
                    ?c1  wdt:P206  wd:Q97.
          MINUS {
                              ?c1  wdt:P206  wd:Q98
                              }
        }
```

- Mostrar entidades ordenadas por su población de manera descendiente.

```
SELECT ?c1 ?c3 WHERE
{
  ?c1  ps:P1082   ?c3.
} ORDER BY DESC( ?c3 )
```

- Mostrar la entidad de mayor población.

```
SELECT ?c1 ?c3 WHERE
{
?c1  ps:P1082 ?c3  .
}ORDER BY DESC( ?c3 )
LIMIT 1
```

3.3. Búsquedas en Wikidata

Para comprobar las características de wikidata haremos un experimento que consistirá en crear árboles de búsqueda sobre algunos de los términos de la inteligencia artificial y nos fijaremos en las características de los árboles de búsqueda. También nos fijaremos en el numero de nodos hoja encontrados.

Se puede ver como seria un árbol de búsqueda en la siguiente imagen:

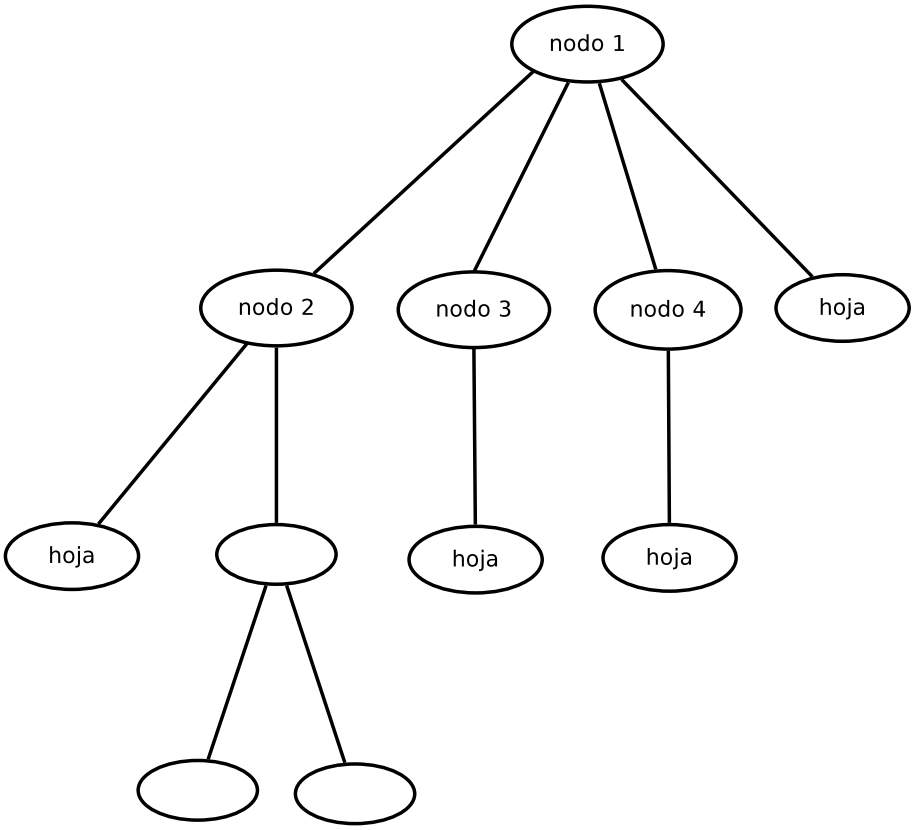


Figura 3.2: Arbol de búsqueda

En este árbol de búsqueda tendría un número total de nodos 11, nodos padres 5. En este árbol tendría un total de 6 hojas. En este árbol el nodo de mayor tamaño sería de un tamaño 4. Haremos un estudio para varios términos distintos ahora vamos a explicar la consulta que realizaremos para comprobar la características de wikidata, si es una red semántica para hacer inferencias o puede tener otros usos.

3.3.1. Estructura de una red semántica: wikidata

Hay varias formas de consultar una red semántica podemos por ejemplo querer buscar información a partir de un nodo buscando los vecinos a partir de ese nodo o podríamos por ejemplo hacer una consulta para encontrar todos los caminos posibles entre dos nodos que serían todas las relaciones existentes entre dos nodos, algo que estaría bien para cuando añadimos nuevas relaciones volver a comprobar las relaciones que existen entre dos nodos puede llevarnos a nuevos descubrimientos.

Nuestra consulta será para obtener información a partir de la red semántica, en la que buscaremos los nodos siguientes a los nodos obtenidos ampliando la información que conocemos sobre el tema del que estamos buscando información la consulta sería de la siguiente forma:

```
select ?q where {
                wd:Q11660   ?a   ?q
                }
```

Donde wd:Q11660 es el nodo del que partimos que si lo buscamos en la wikipedia es el término correspondiente a inteligencia artificial.

A partir de aquí ejecutamos una consulta recursiva a la base de datos en la que se encadena varias consultas hasta que creamos un árbol de búsqueda en el que se puede ver la información del nodo, podemos escoger las profundidad del árbol de búsqueda y también nos fijaremos en parámetros del árbol de búsqueda como el máximo tamaño del nodo, el número total de nodos y el número total de hojas.

3.3.2. Algoritmo

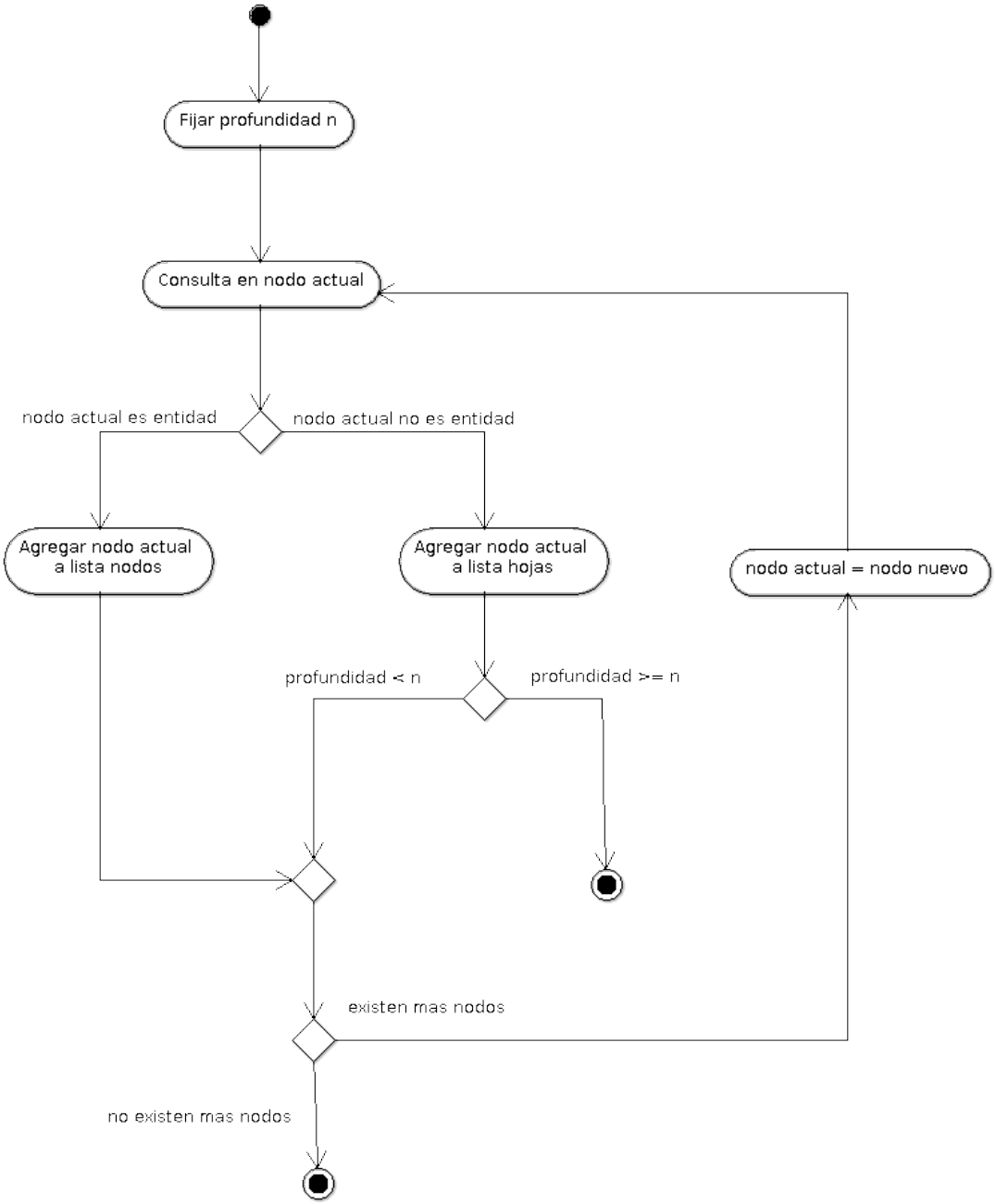


Figura 3.3: Algoritmo

3.3.3. Realización del experimento

Los términos que utilizaremos para el experimento tendrán que ver con el contexto de la informática.

Arbol de búsqueda palabra internet

La palabra internet en wikidata se conoce con la identidad wd:Q8366.

Realizamos la consulta para visualizar las características del árbol de búsqueda con distintas profundidades y obtenemos los siguientes resultados:

```
profundidad del arbol:3
numero total nodos:18
numero total hojas:70
max node size:33
```

```
profundidad del arbol:4
numero total nodos:2008
numero total hojas:7679
max node size:890
```

```
profundidad del arbol:5
numero total nodos:17958
numero total hojas:70143
max node size:970
```

Realizaremos una muestra para árboles de búsqueda de altura 3 y otra para árboles de altura 4, para altura 5 sería ya mucho tiempo para cada observación.

3.3.4. Datos árbol búsqueda altura 3

Arbol de búsqueda altura 3			
Unidades	**Nodos**	**Hojas**	**Nodo maximo tamaño**
wd:Q68	361	1840	414
wd:Q6723676	124	224	131
wd:Q39645	205	945	335
wd:Q4055684	82	354	215
wd:Q408386	83	288	140
wd:Q1156402	92	273	128
wd:Q1301371	158	839	442
wd:Q812527	53	465	361
wd:Q5862903	423	1042	408
wd:Q7397	1582	5543	630
wd:Q224821	134	448	215
wd:Q201413	162	675	224
wd:Q1350299	10	33	34
wd:Q817862	63	198	67
wd:Q1415372	41	214	194
wd:Q1128326	34	131	103
wd:Q177719	203	751	192
wd:Q8269924	200	1347	597
wd:Q1151406	21	60	62
wd:Q223655	198	778	200
wd:Q309100	62	378	138
wd:Q151885	142	736	239
wd:Q43054	473	1718	359
wd:Q451968	37	469	361
wd:Q10388919	1	7	7
wd:Q315	641	2541	399
wd:Q609057	20	112	91
wd:Q163468	274	875	330
wd:Q1051925	56	270	192
wd:Q177929	38	207	126
wd:Q837528	37	137	73
wd:Q1661153	4	11	11
wd:Q3707858	72	313	116
wd:Q212542	131	1087	383
wd:Q1921838	7	13	14
wd:Q11538	83	591	227
wd:Q3123740	58	279	144
wd:Q11023	283	1253	345
wd:Q466	597	2235	581
wd:Q36484	98	617	225
wd:Q1664689	62	487	383

En esta tabla se pueden ver las entidades de wikidata que son parte de la muestra para árboles de altura 3 las hemos escogido en base al contexto de la informática ya que es un contexto extendido en la wikipedia y en internet y que puede ser válido para realizar el experimento.

3.3.5. Gráficas frecuencia altura 3

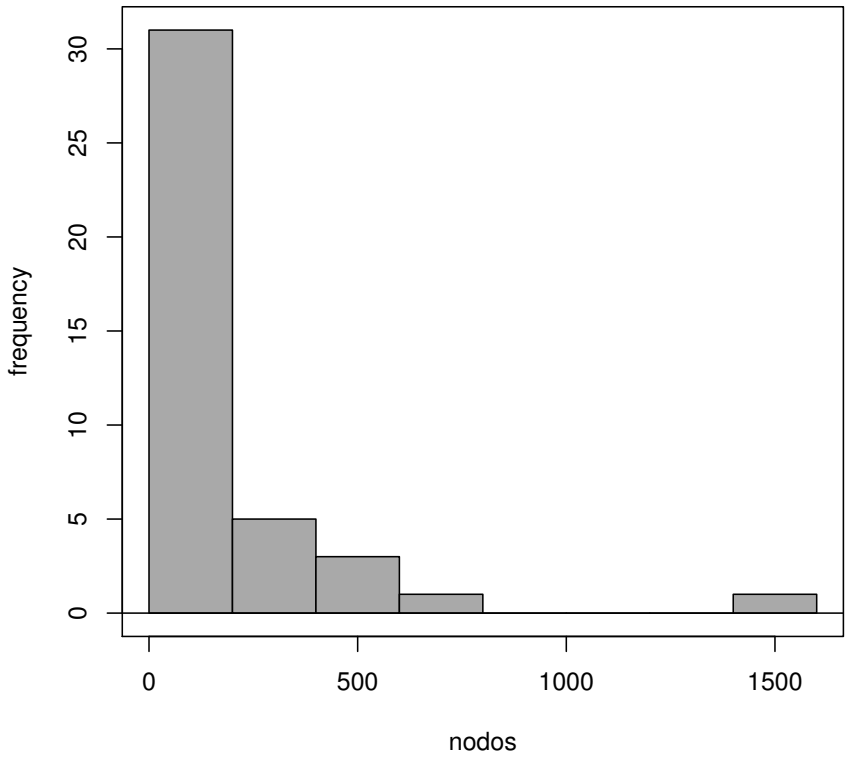


Figura 3.4: Histograma nodos

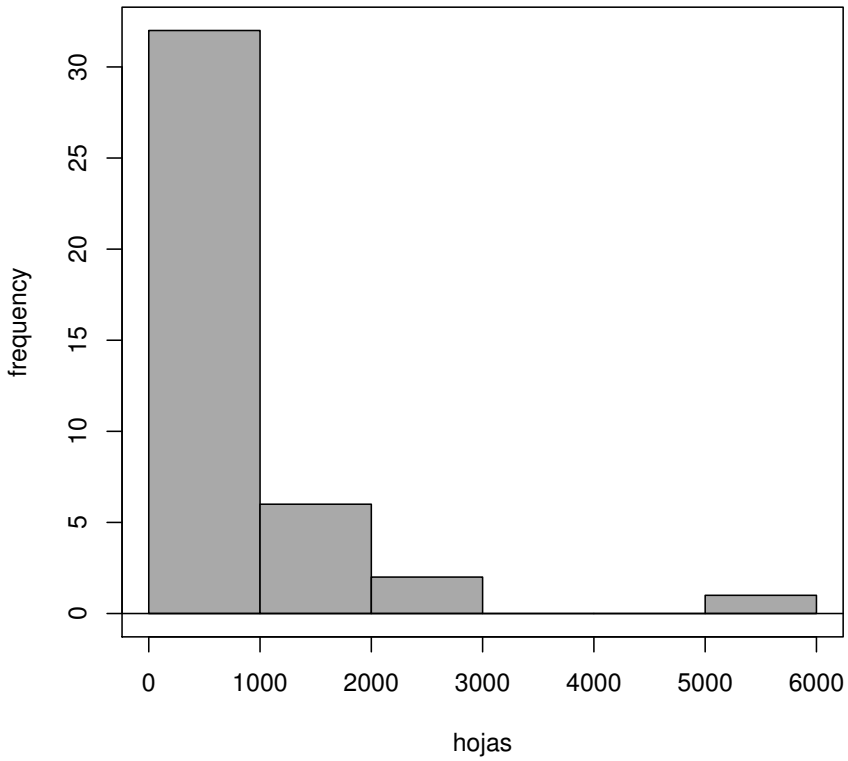


Figura 3.5: Histograma hojas

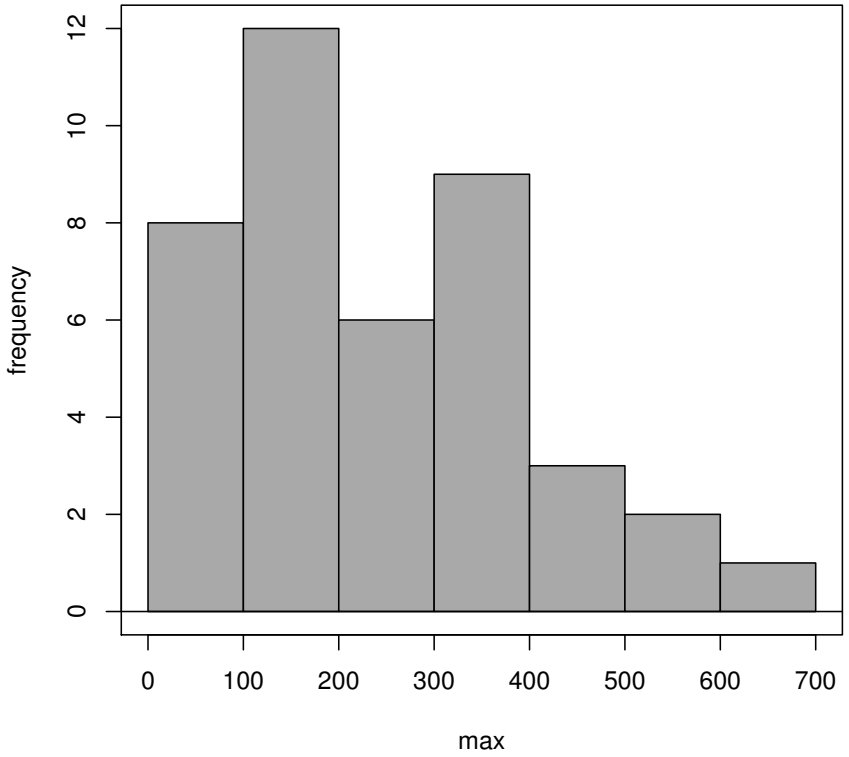


Figura 3.6: Histograma max nodo

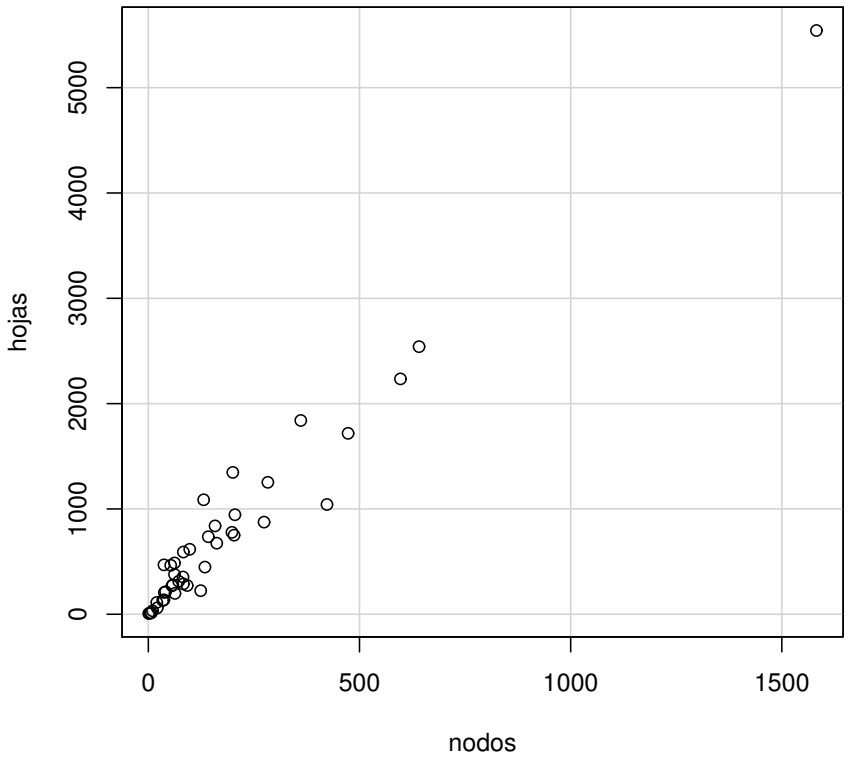


Figura 3.7: Dispersión hojas nodos

Se puede ver como la dispersión hojas nodo se mantiene aproximadamente creando una diagonal lo que indica que el arbol de búsqueda tiene una relación de hojas nodos similar para toda la muestra.

3.3.6. Gráficas altura 4

Arbol de búsqueda altura 4			
Unidades	**Nodos**	**Hojas**	**Nodo máximo tamaño**
wd:Q68	2929	13047	630
wd:Q6723676	2325	3409	1550
wd:Q39645	1541	7160	399
wd:Q4055684	1769	3506	860
wd:Q408386	432	2332	384
wd:Q1156402	454	1724	196
wd:Q1301371	672	3796	442
wd:Q812527	597	3947	658
wd:Q5862903	6821	12751	1061
wd:Q7397	17613	51154	667
wd:Q224821	703	3056	411
wd:Q201413	1119	4636	411
wd:Q1350299	12	34	34
wd:Q817862	430	1401	352
wd:Q1415372	438	2004	411
wd:Q1128326	131	648	114
wd:Q177719	2560	7000	910
wd:Q8269924	1507	8957	599
wd:Q1151406	26	61	62
wd:Q223655	1493	6894	384
wd:Q309100	268	1651	384
wd:Q151885	1970	4925	910
wd:Q43054	3192	9955	384
wd:Q451968	500	3280	667
wd:Q10388919	1	7	7
wd:Q315	6487	27263	631
wd:Q609057	60	585	384
wd:Q163468	4151	12976	1805
wd:Q1051925	279	1575	252
wd:Q177929	105	578	175
wd:Q837528	158	521	117
wd:Q1661153	4	11	11
wd:Q3707858	564	2813	599
wd:Q212542	567	3721	384
wd:Q1921838	9	14	14
wd:Q11538	400	2331	384
wd:Q3123740	1597	2978	918
wd:Q11023	1884	9165	384
wd:Q466	7770	25433	1434
wd:Q36484	519	2944	384
wd:Q1664689	297	1760	383

Ahora podemos observar las variables para un árbol de altura 4 son los mismos conceptos que para el árbol de altura 3.

3.3.7. Gráficas árbol altura 4

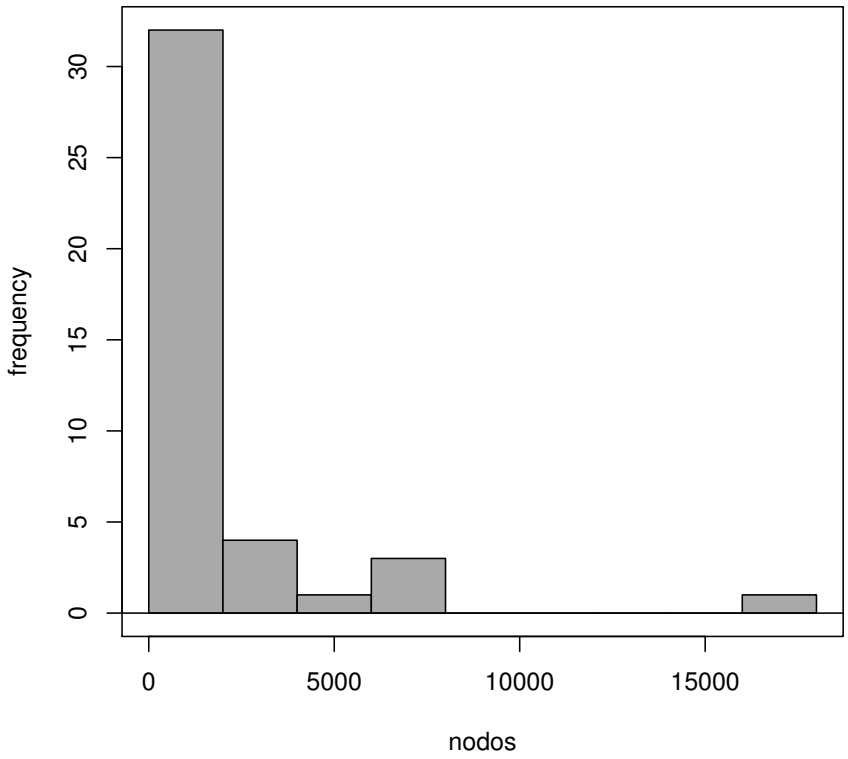


Figura 3.8: Histograma nodos

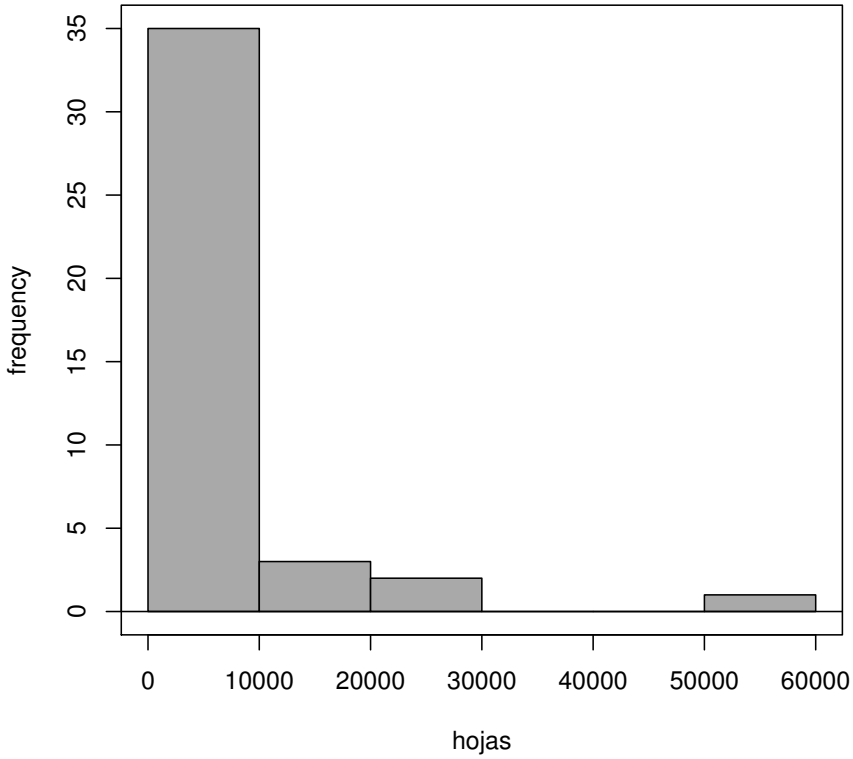


Figura 3.9: Histogramas hojas

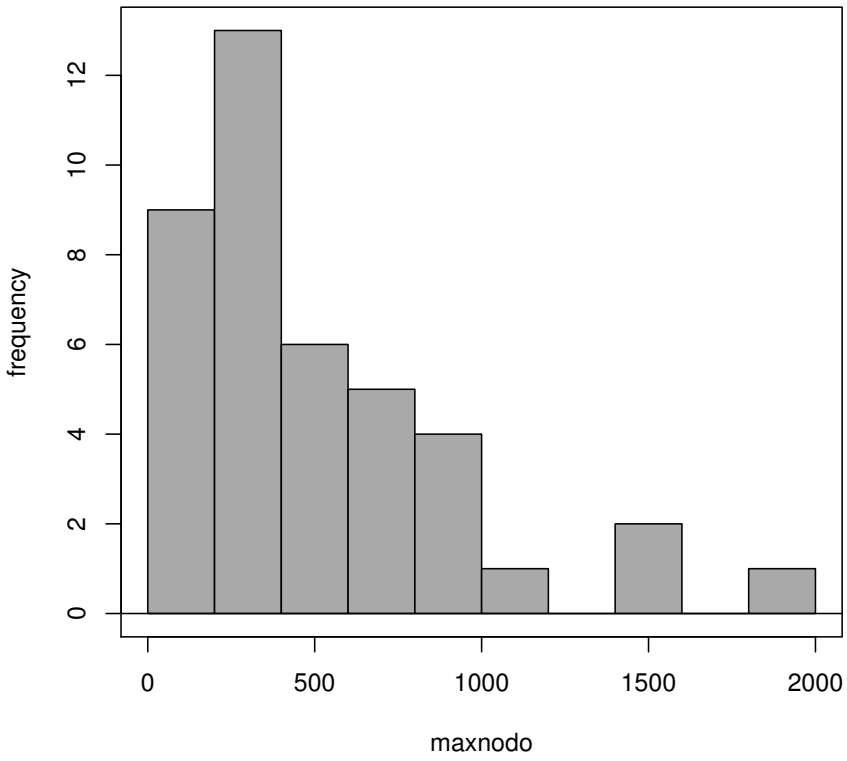


Figura 3.10: Histograma max nodo

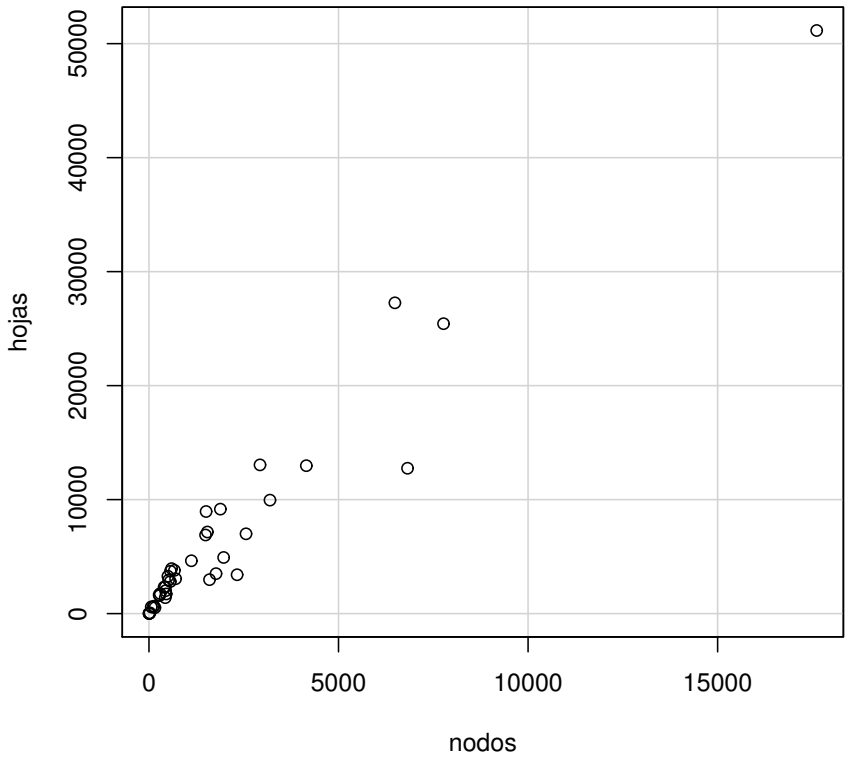


Figura 3.11: Dispersión hojas nodos

En este caso aunque todavía se puede intuir una diagonal ya se puede ver como empieza a tener más diferencias en la relación nodos hojas.

3.4. Servicios web

Los servicios web se utilizan para publicar funcionalidades a través de internet que permiten a los usuarios usar la funcionalidad de un software a distancia.

Wikidata consta de un servicio web en el que se puede realizar consultas y obtener resultados.

```
public String obtenerResultadoSparql(String consulta){
        URL url=null;
        try {
                url = new URL("https://query.wikidata.org/sparql?query="
        +URLEncoder.encode(consulta, "UTF-8"));
        } catch (MalformedURLException e) {
                e.printStackTrace();
        } catch (UnsupportedEncodingException  e) {
                e.printStackTrace();
        }


        HttpURLConnection connection = null;
    try {
                connection = (HttpURLConnection)url.openConnection();
        } catch (IOException e) {
                e.printStackTrace();
        }

    try {
                connection.setRequestMethod("GET");
        } catch (ProtocolException e1) {
                e1.printStackTrace();
        }

    connection.setRequestProperty("Accept", "application/sparql-results+xml");

    try {
                if(connection.getResponseCode() != 200){
                  System.out.println("fallo en la conexion");
                 }
        } catch (IOException e) {
                e.printStackTrace();
        }

    BufferedReader reader = null;
     try {
                reader = new BufferedReader(
                                new InputStreamReader(connection.getInputStream()))
        } catch (IOException e) {
                e.printStackTrace();
        }

     String result="";
     String output="";
     try {
                  while((output = reader.readLine()) != null){
```

```
                        result += output + "\n";
                    }
            } catch (IOException e) {
                    e.printStackTrace();
            }

        return result;
}
```

3.5. Estructura documentos ttl de wikidata

La información en wikidata se encuentra disponible en ficheros ttl.

Cuando buscamos información en wikidata nos encontramos con las uris de las entidades por ejemplo la uri acerca de una entidad `https://www.wikidata.org/wiki/Q42` y podríamos acceder a la web de la entidad, pero si queremos acceder al fichero ttl tendriamos que utilizar la siguiente uri por ejemplo `http://www.wikidata.org/wiki/Special:EntityData/Q42.ttl` en este fichero encontramos toda la información semántica de la entidad.

Los prefijos se encuentran al principio del fichero:

Observamos los prefijos en el fichero:

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix wikibase: <http://wikiba.se/ontology-beta#> .
@prefix wds: <http://www.wikidata.org/entity/statement/> .
@prefix wdata: <https://www.wikidata.org/wiki/Special:EntityData/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix schema: <http://schema.org/> .
@prefix cc: <http://creativecommons.org/ns#> .
@prefix geo: <http://www.opengis.net/ont/geosparql#> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix wdref: <http://www.wikidata.org/reference/> .
```

Estos prefijos se usan en el documento ttl para manejar más adecuadamente las uris y hacer el código del documento y las consultas sparql más legibles.

También al principio del fichero después de los prefijos nos encontramos con un resumen de la información que contiene el fichero:

```
wdata:Q42 a schema:Dataset ;
        schema:about wd:Q42 ;
        cc:license <http://creativecommons.org/publicdomain/zero/1.0/> ;
        schema:softwareVersion "0.1.0" ;
        schema:version "621144780"^^xsd:integer ;
        schema:dateModified "2018-01-16T21:39:15Z"^^xsd:dateTime ;
        wikibase:statements "133"^^xsd:integer ;
        wikibase:identifiers "73"^^xsd:integer ;
        wikibase:sitelinks "102"^^xsd:integer .
```

En este caso podemos ver la información acerca de la entidad Q42 que tiene 133 wikibase:statements, 73 wikibase:identifiers y 102 wikibase:sitelinks.

En el siguiente trozo del fichero ttl podemos ver las propiedades con sus etiquetas:

```
wd:P31 a wikibase:Property ;
        rdfs:label "instance of"@en ;
        skos:prefLabel "instance of"@en ;
        schema:name "instance of"@en ;
        rdfs:label "nature de l'élément"@fr ;
        skos:prefLabel "nature de l'élément"@fr ;
        schema:name "nature de l'élément"@fr ;
        rdfs:label "ist ein(e)"@de ;
        skos:prefLabel "ist ein(e)"@de ;
        schema:name "ist ein(e)"@de ;
        rdfs:label "istanza di"@it ;
        skos:prefLabel "istanza di"@it ;
```

Podemos ver que las etiquetas se muestran en muchos idiomas distintos.

Resto de etiquetas relativas a propiedades

```
a wikibase:Property ;
        wikibase:propertyType <http://wikiba.se/ontology-beta#WikibaseItem
        wikibase:directClaim wdt:P31 ;
        wikibase:claim p:P31 ;
        wikibase:statementProperty ps:P31 ;
        wikibase:statementValue psv:P31 ;
        wikibase:qualifier pq:P31 ;
        wikibase:qualifierValue pqv:P31 ;
        wikibase:reference pr:P31 ;
        wikibase:referenceValue prv:P31 ;
        wikibase:novalue wdno:P31 .
```

Aquí al acabarse las etiquetas en varios idiomas comienza la información sobre las relaciones entre nodos de wikidata, se pueden ver las etiquetas de Douglas adams y luego el resto de relaciones en sujeto, predicado y objeto.

```
            "Douglas Noel Adams"@tr,
            "Douglas N. Adams"@tr,
            "Douglas NoÃ«l Adams"@et ;
    wdt:P31 wd:Q5 ;
    wdt:P21 wd:Q6581097 ;
    wdt:P106 wd:Q214917,
            wd:Q28389,
            wd:Q6625963,
            wd:Q4853732,
            wd:Q18844224,
            wd:Q245068,
            wd:Q487596 ;
    wdt:P800 wd:Q25169,
            wd:Q902712,
            wd:Q7758404,
            wd:Q578895,
            wd:Q721,
            wd:Q1042294,
            wd:Q187655 ;
    wdt:P569 "1952-03-11T00:00:00Z"^^xsd:dateTime ;
    wdt:P19 wd:Q350 ;
    wdt:P570 "2001-05-11T00:00:00Z"^^xsd:dateTime ;
    wdt:P1196 wd:Q3739104 ;
    wdt:P509 wd:Q12152 ;
    wdt:P20 wd:Q159288 ;
```

3.6. Extraer información de wikidata

Quizás nos sea útil realizar algunas consultas para extraer información de wikidata, consultando los ficheros ttl que nos podemos descargar de wikidata y seleccionando aquella información que nos sea útil, obteniendo así un nuevo documento con la información que nosotros creamos oportuna y presentarla en forma de web y guardarla también en ttl para que se pueda almacenar en otros lugares para su reutilización.

Con la siguiente consulta extraemos toda las relaciones que tienen a la entidad Q42 como nodo a la izquierda.

```
select ?p ?q where {
        <http://www.wikidata.org/entity/Q42> ?p ?q
        }
```

Capítulo 4

Un entorno local de apoyo

También hemos desarrollado una pequeña web que nos permite trabajar con documentos semánticos, donde se pueden subir documentos semánticos, consultarlos y explorarlos. La página web esta hecha en JEE utilizando las herramientas que existen para el desarrollo de páginas web con contenidos semánticos como la librería Jena que permite consultar ficheros con el lenguaje Sparql y queda un aspecto similar a un endpoint Sparql. Para crear la web y que se pudieran almacenar información en el servidor hemos utilizado la librería arq de apache jena, que nos permitía almacenar y consultar los ficheros, también hemos utilizado una base de datos donde se almacena y gestiona toda la información de los ficheros.

El objetivo es un sistema en el que crear y almacenar para propósitos educativos documentos semánticos el sistema consta de dos tipos de usuarios, profesores y administradores, los profesores podrán consultar y crear documentos semánticos dentro de los grupos que asigne el administrador, los profesores estan agrupados por asignaturas, por grupos de investigación y por departamentos.

4.1. Login

La página de login permite gestionar la página como administrador o utilizar la página como profesor.

Figura 4.1: Página login

4.2. Menú administrador

El menú de administrador desde el que se puede acceder a gestionar la web.

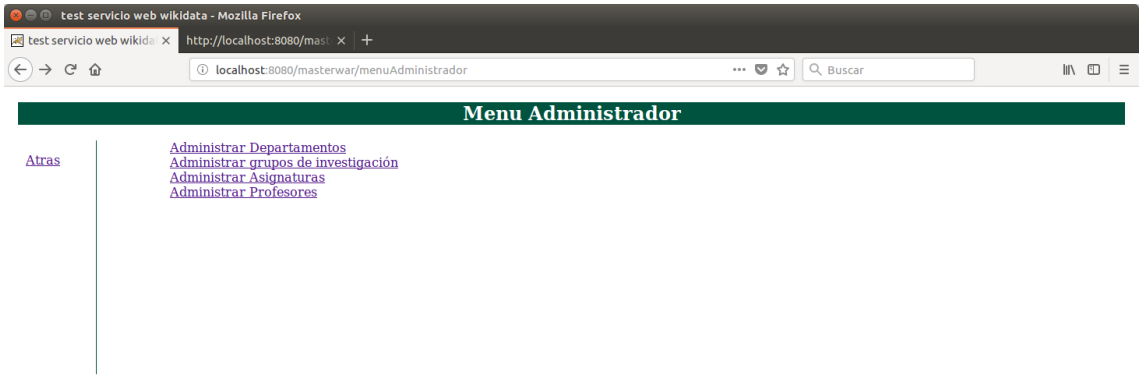


Figura 4.2: Menú administrador

4.3. Gestión de Profesores

Se utiliza para gestionar los profesores por ejemplo crear nuevos profesores o eliminarlos.

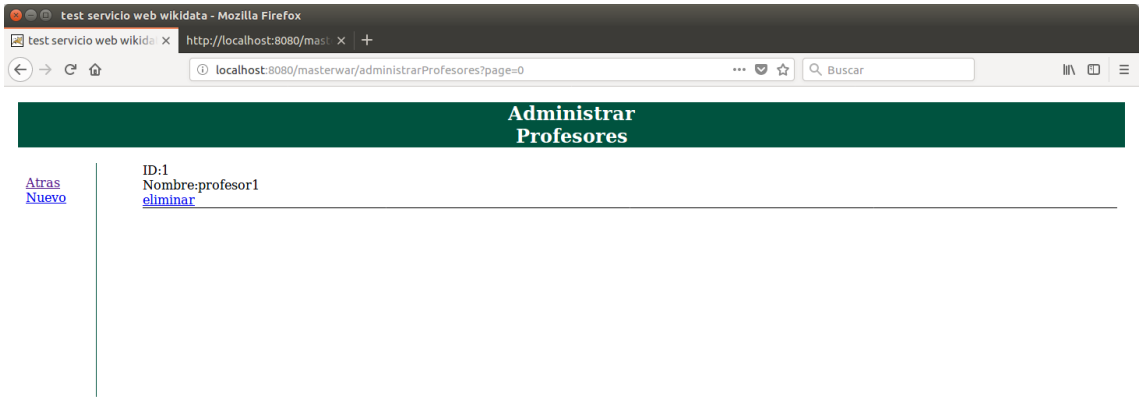


Figura 4.3: Gestión profesores

4.4. Creación de la asignatura, departamentos o grupo de investigación

El administrador crea la asignatura u otro grupo en la sección de la página web habilitada para tal función y también decide que profesores pueden ver y editar los documentos de la asignatura.

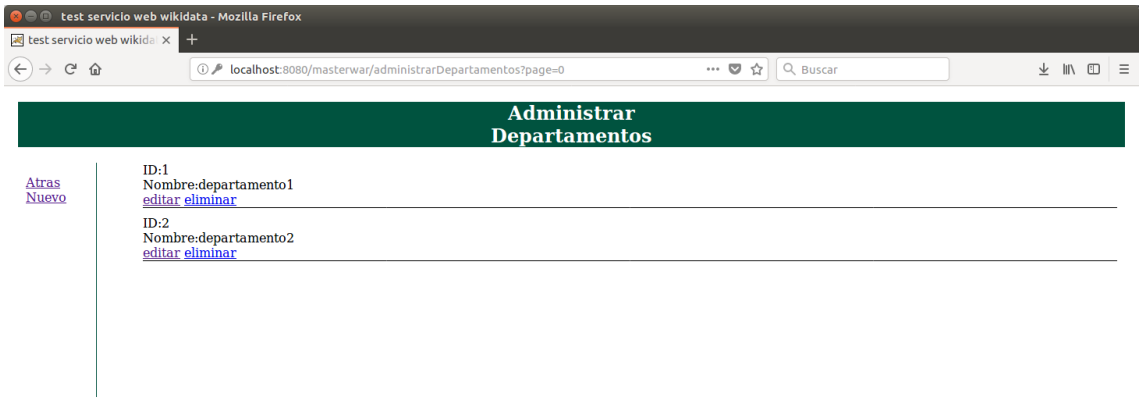


Figura 4.4: Administrar departamentos

4.5. Edición de los departamentos por parte del profesor

En editar departamento se escogen los profesores que podrán editar los documentos se eligen agrega ándolos a la lista inferior.

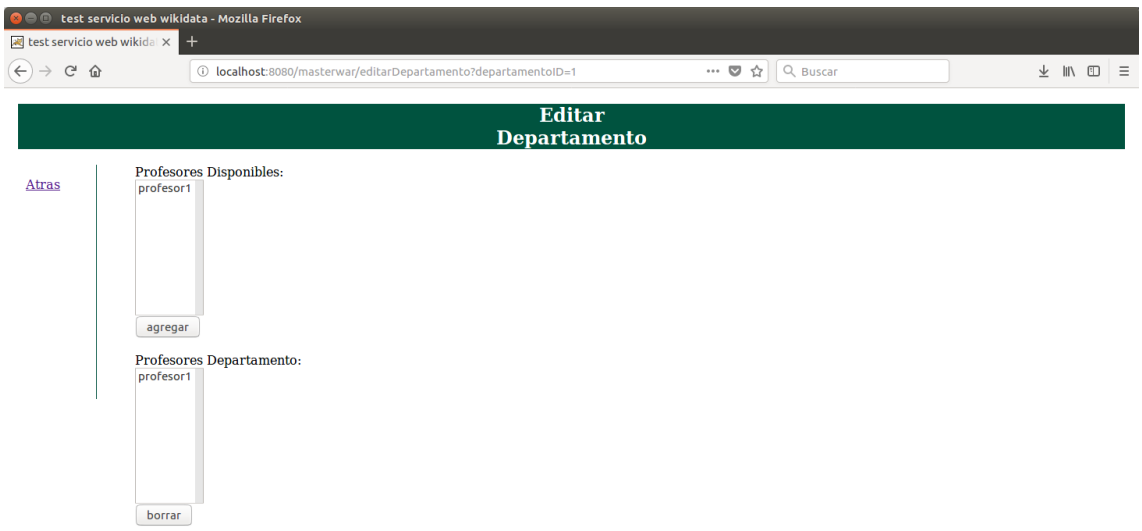


Figura 4.5: Editar departamento

4.6. Gestor departamentos

Donde se gestionan los departamentos del profesor.
Tiene la opción de ver los documentos de los departamentos.

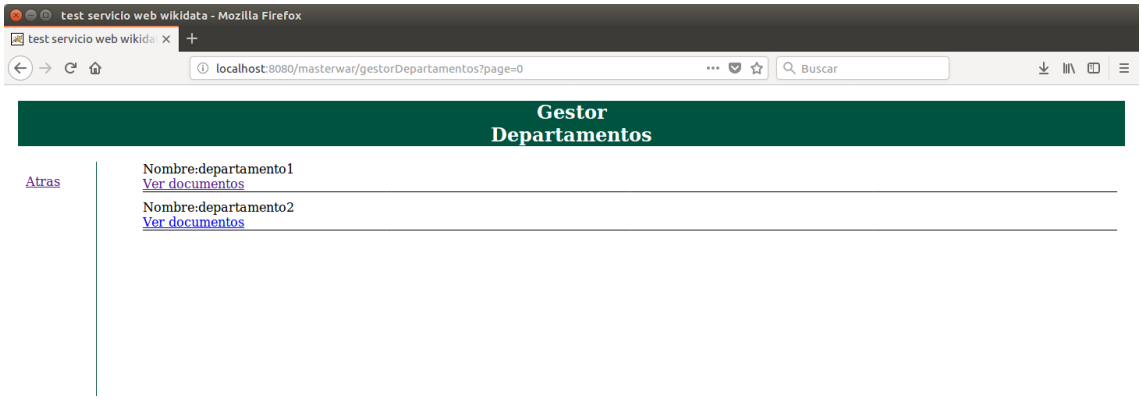


Figura 4.6: Gestor departamentos

4.7. Creación de carpetas compartidas

Existe un grupo especial que es el de las carpetas compartidas que puede ser creadas sin necesidad de ser administrador.

Figura 4.7: Nueva carpeta compartida

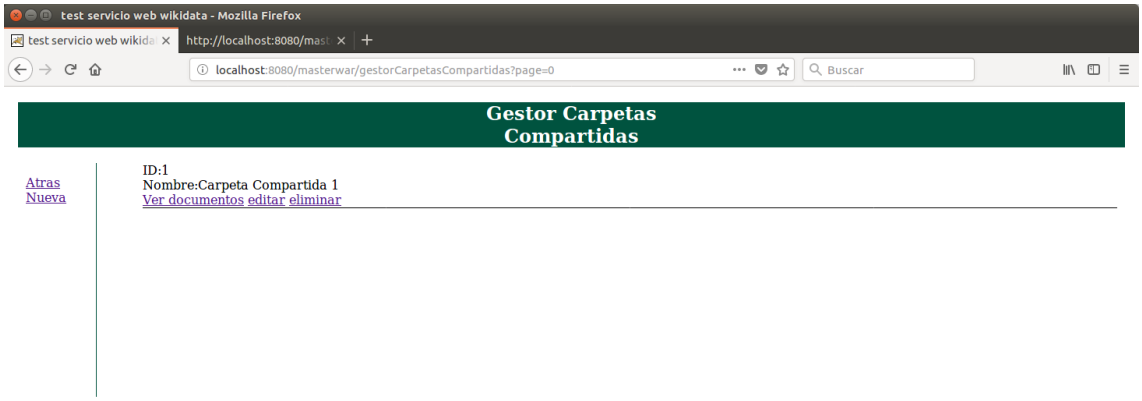


Figura 4.8: Gestor carpetas Compartidas

4.8. Operaciones con los ficheros

En la opción de visualizar ficheros se pueden visualizar todos los ficheros y debajo del nombre de cada fichero las operaciones que se pueden realizar con ellos.

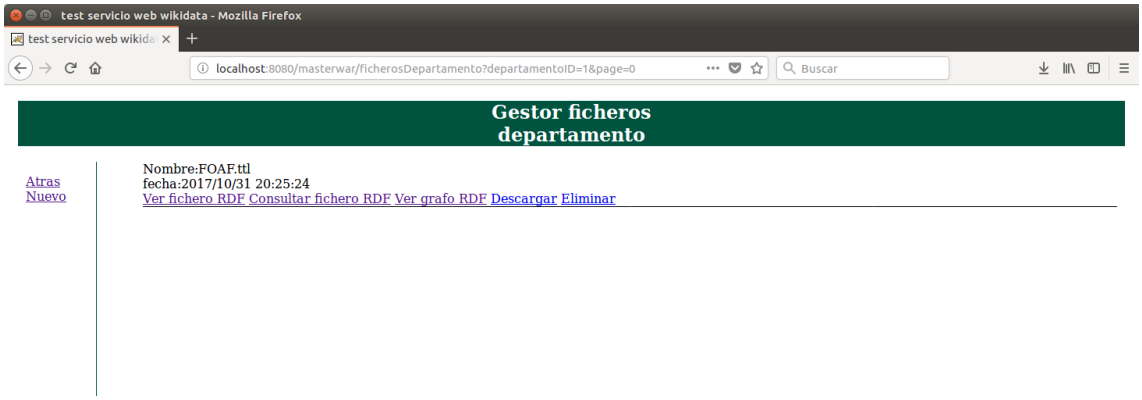


Figura 4.9: Gestor ficheros

4.9. Visualizar ficheros semánticos en la web

Entre las opciones de las operaciones que se pueden realizar con los ficheros está la de visualizar el fichero tal cual esta escrito en cualquiera de los formatos RDF, por ejemplo el siguiente esta en .ttl.

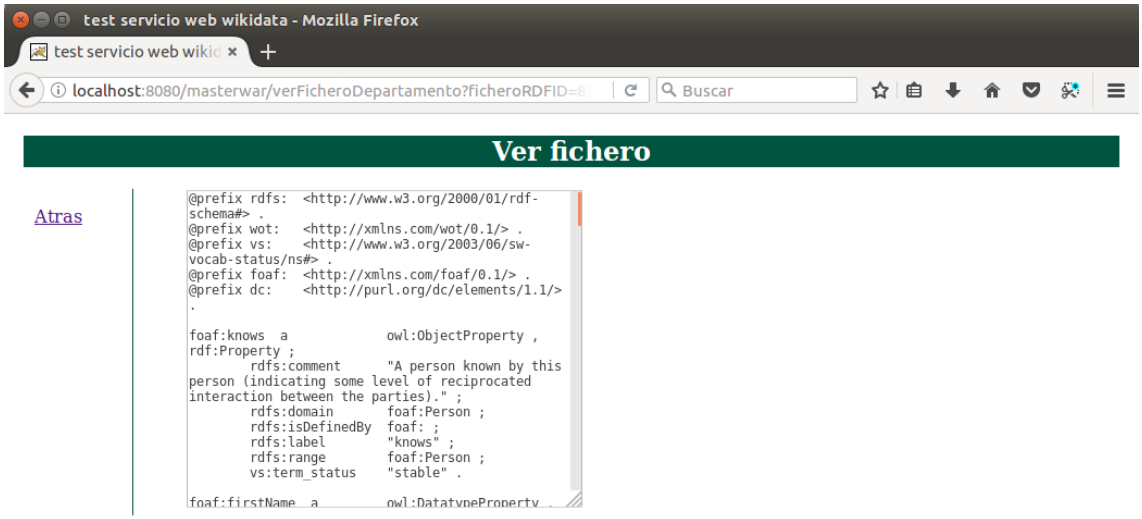


Figura 4.10: Ver fichero

4.10. Consulta de ficheros en Sparql

También se pueden consultar los ficheros como si se tratase de un endpoint, primero realizando la consulta y luego visualizando el resultado en xml, aunque jena permite ver el resultado en diversos formatos.

Figura 4.11: Consulta

Figura 4.12: Resultado

4.11. Exploración de ficheros

También se pueden explorar los ficheros semánticos mediante un visor creado especificamente para visualizar las relaciones entre nodos y aristas de los ficheros rdf donde se pueden contemplar la traza de varios nodos y aristas relacionados.

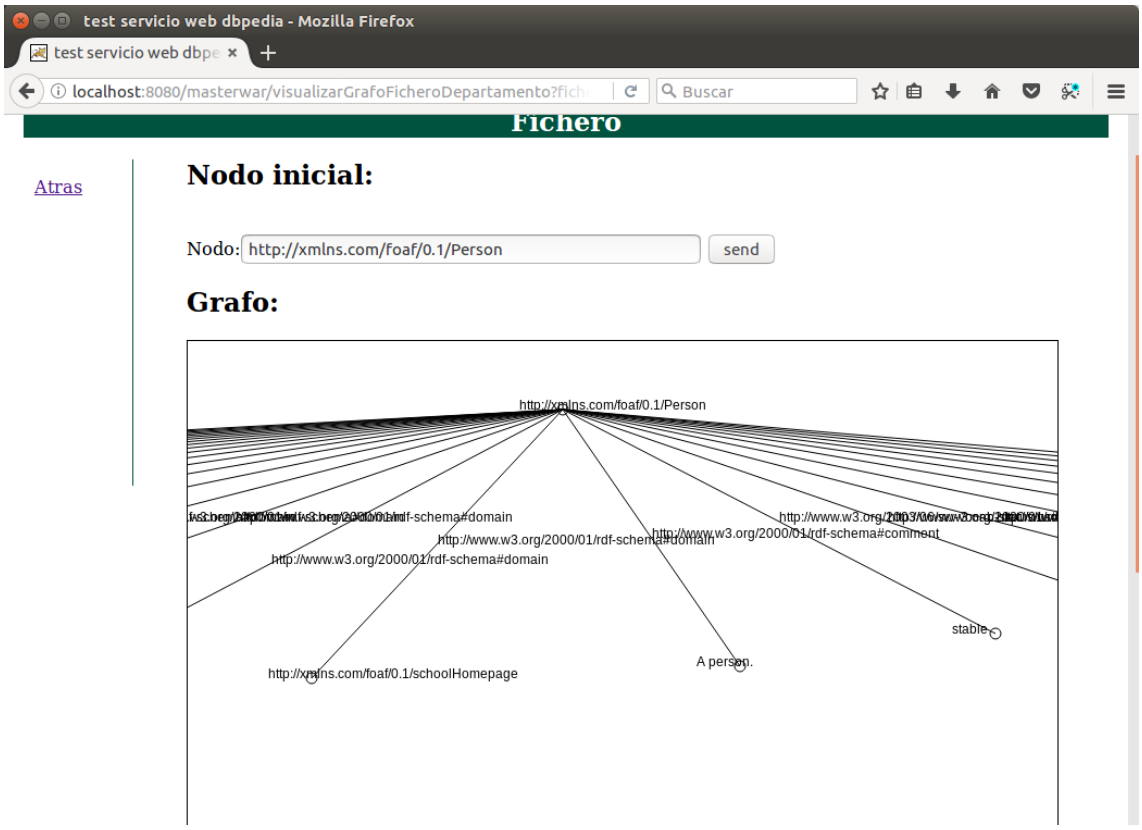


Figura 4.13: Explorar fichero web

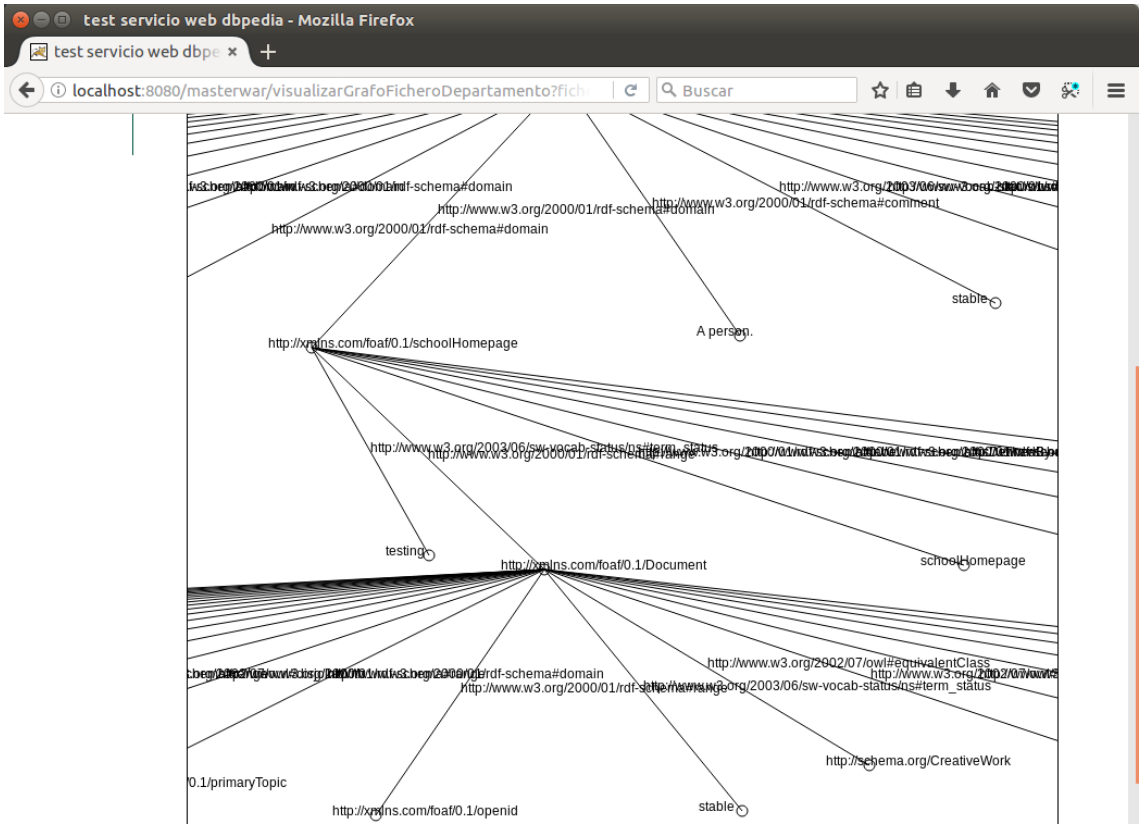


Figura 4.14: Explorar fichero web extendido

4.12. Integración de la web con Wikidata

Hemos creado una sección en la web que nos permite explorar wikidata, y extraer las principales relaciones de las páginas de wikidata, también hemos añadido la funcionalidad de añadir cualquier relación semántica que nos encontremos en las páginas de wikidata a un fichero obteniendo un fichero con las relaciones semánticas que luego podremos incorporar a los ficheros rdf.

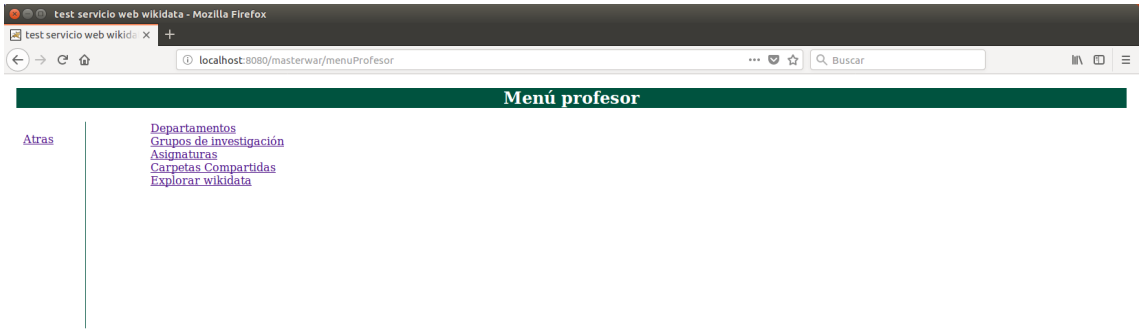


Figura 4.15: Menú explorar wikidata

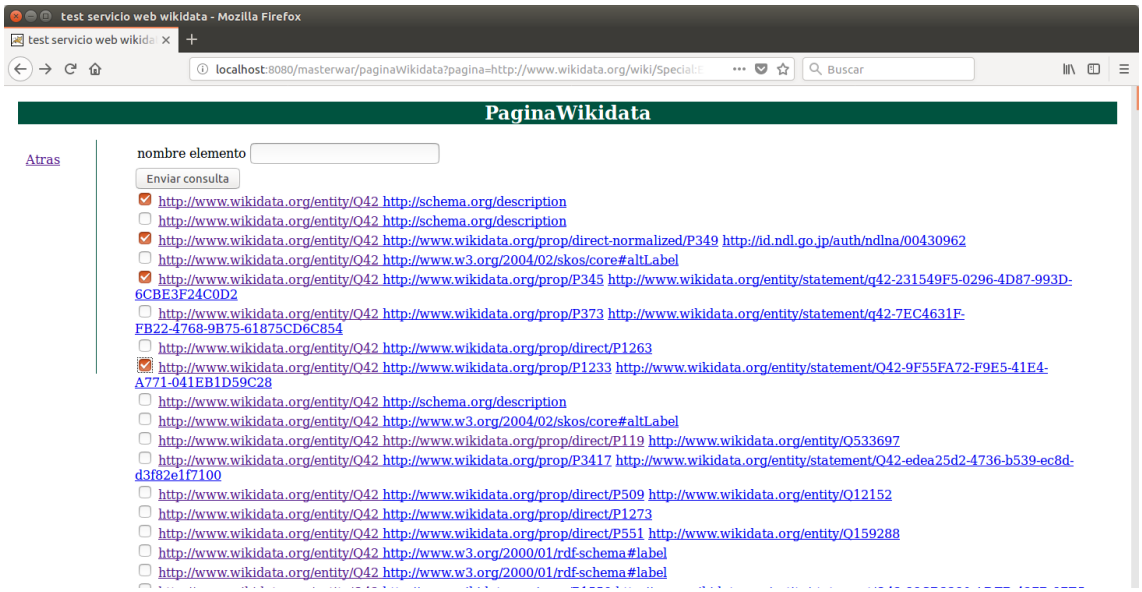


Figura 4.16: Selección de relaciones wikidata

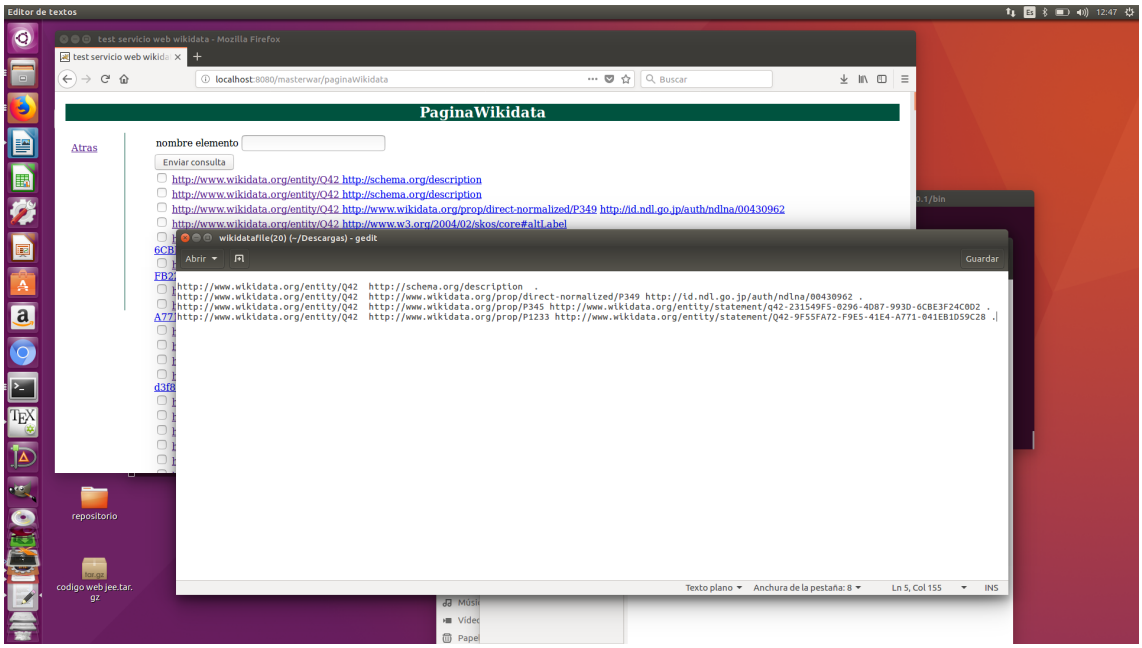


Figura 4.17: Fichero resultado

4.13. Características técnicas de la pagina web

La página web esta realizada con una arquitectura de 3 capas diferenciadas una capa de interfaz, una de lógica, y otra de persistencia de los datos.

La capa de interfaz consta de todos los servlets que se utilizan para crear la interfaz y conseguir que esta sea interactiva.

La capa de lógica se encarga de gestionar la lógica de la página web, operaciones que pueden realizar administradores y profesores, y operaciones que se pueden realizar sobre departamentos asignaturas y grupos de investigación.

La capa de persistencia esta hecha en java también utilizando una base de datos SQL y el patrón DAO para representar los objetos de la aplicación web en la base de datos.

Características importantes de la página web son el uso de sesión mediante cookies de sesión, tanto para administrador como para usuario. En un futuro se podría mejorar la web mediante el uso de SSL para las comunicaciones seguras.

4.14. Tecnologías utilizadas

4.14.1. Eclipse

Eclipse es un entorno de desarrollo que asiste a desarrollo de software en múltiples lenguajes de programación, hemos utilizado eclipse por que nos permite desarrollar en el lenguaje de programación java que es indicado para el desarrollo de páginas web utilizando las tecnologías de JEE.

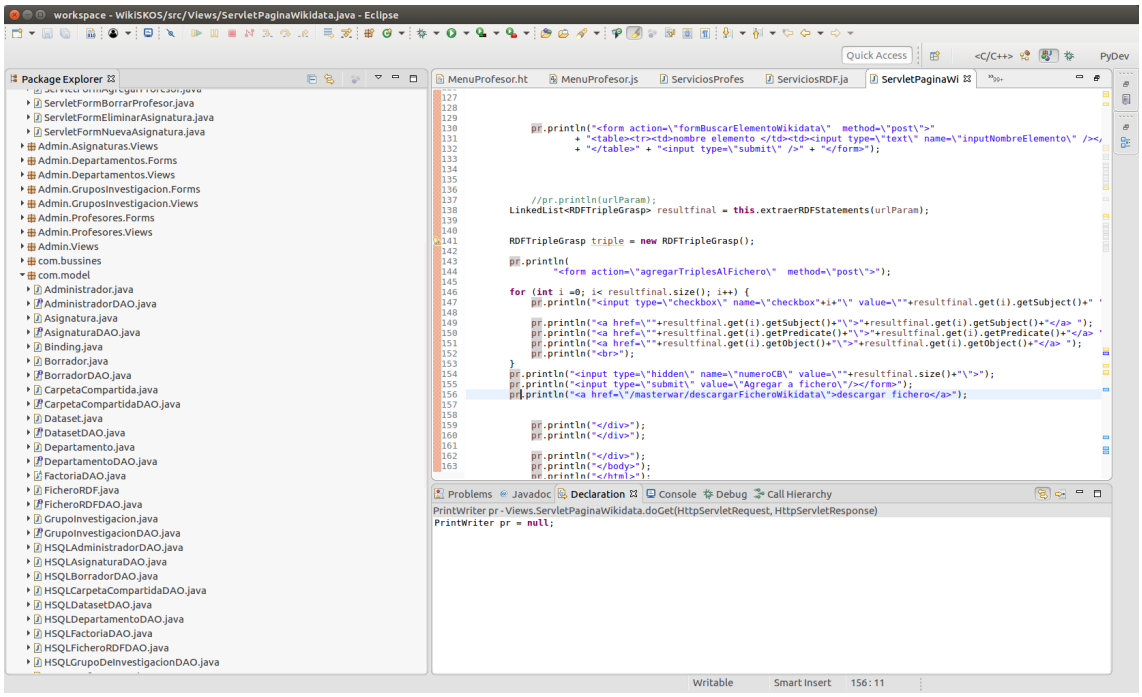


Figura 4.18: Eclipse

4.14.2. HSQL

Es un sistema de bases de datos sencillo que hemos utilizado para el desarrollo de la página web, permite la creación de tablas sql para la implementación de esquemas relacionales en sql.

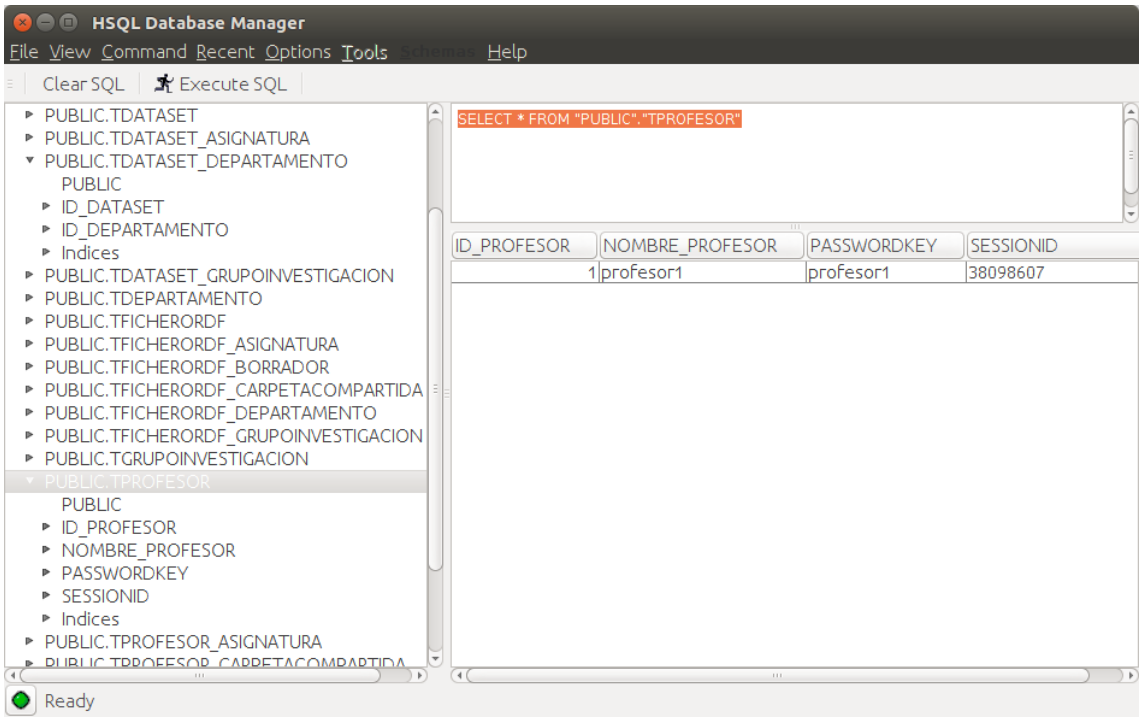


Figura 4.19: HSQL

4.14.3. Servidor de aplicaciones Apache Tomcat

El servidor de aplicaciones es uno de los componentes de JEE más importantes y que se encarga de que la aplicación web este disponible para utilizarla a través de internet, tomcat es uno de los servidores de aplicaciones más habituales, ha sido desarrollado por apache y para el se han realizado numerosas implementaciones de componentes de JEE.

La instalación de la aplicación en el servidor web consiste en una aplicación web contenida en un archivo .war

4.14.4. ANT

Apache ant es un sistema utilizado para gestionar la compilación de las aplicaciones realizadas en java en nuestro caso lo utilizamos cada vez que realizamos cambios en algunos de los componentes de la web.

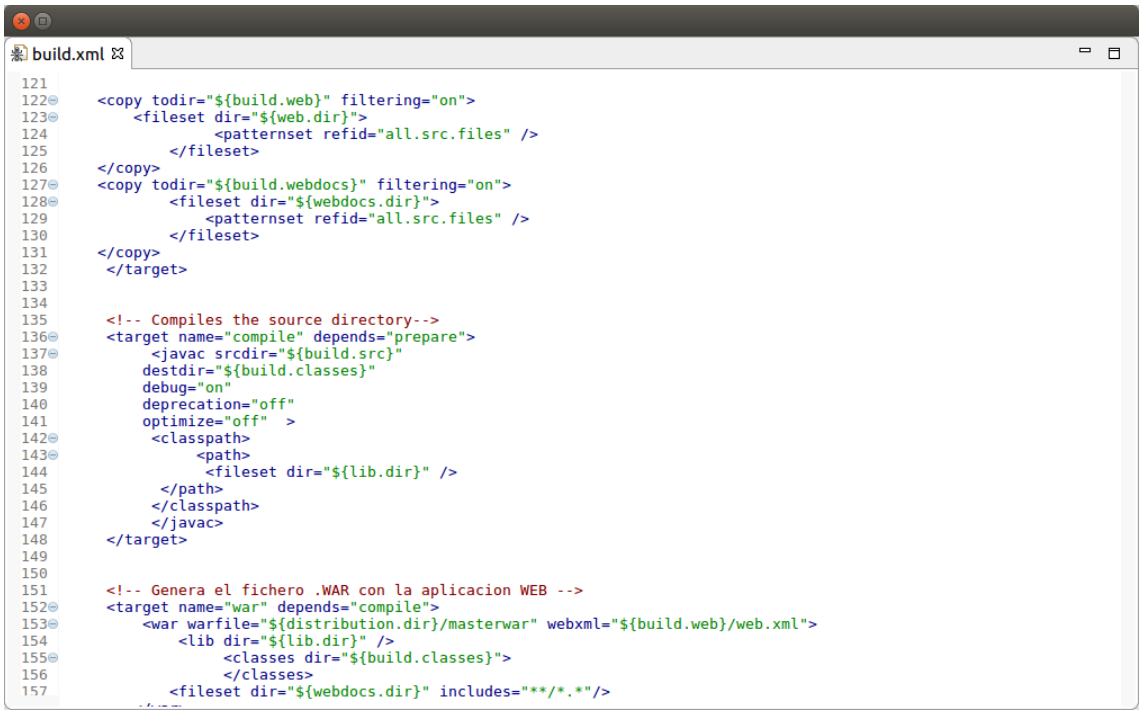


Figura 4.20: Ant

4.14.5. Javascript

Es un lenguaje de programación que utilizan los desarrolladores cuando quieren ejecutar código de una aplicación web en el navegador utilizando el ordenador del usuario, nosotros lo utilizamos para realizar algún programa que nos permite hacer algún dibujo y visualizaciones especiales durante el uso de la aplicación web.

4.14.6. CSS

Es un lenguaje para la definición de la interfaz de las páginas web.

Capítulo 5

Conclusiones

Se pretendía confirmar que el contexto tecnológico permitía ya la aplicación buscada: recabar gran número de datos enlazados para su uso en actividades formativas. Sin valorar otros datasets, sólo en Wikidata, se ha confirmado suficiente el número de tripletas y la riqueza semántica de sus clases y propiedades; en este caso, para usos diversos. La consulta a otros datasets existentes, masivos especializados (p.ej. sobre estadísticas económicas), permitiría lo mismo para asignaturas más especializadas.

Sobre este conocimiento se han desarrollado dos estrategias de consulta, siempre pensando en convertirlas en plantillas o en procesos iterativos para un usuario no técnico. La estrategia más general desarrollada tiene tres pasos: (1) admite una etiqueta o una primera búsqueda que facilita un conjunto de ítem y sus propiedades, (2) hay un proceso de refinamiento de ternas asociados a los ítem y (3) sobre todos los ítem del conjunto resultante se vuelve a disparar una consulta ampliatoria a Wikidata. Aparte de esta estrategia general se han probado otras plantillas de preguntas más dirigidas, en alguno de los ejes detectado en el análisis del modelo de Wikidata: en un determinado contexto se pregunta sólo por datos temporales, o por datos geoespaciales, o por subpropiedades o subclases relacionadas con la anterior.

En cuanto a los resultados de estas preguntas, sobre Wikidata, se puede obtener como resultado un conjunto general de recursos propios textuales o enlaces externos desde estos ítem: p.ej. identificadores del ítem en otros datasets externos, archivos multimedia en WikiCommons o un volcado de las páginas de Wikipedia relacionadas con el ítem en los idiomas que se escoja.

Bibliografía

[1] B. Carpenter. *The Logic of Typed Feature Structures*. 2005.

[2] J. H. Daniel Jurafsky. *Speech and Language Processing*. 2 edition, 2009.

[3] M. Fitting. *First-Order Logic and Automated Theorem Proving*. 1996.

[4] D. C. Franz Baader. *The Description Logic Handbook*. 2010.

[5] D. B. Frederick Hayes-Roth, Donald A.Waterman. *Building Expert Systems*. 1983.

[6] J. F. Patrice Boizumault, Ara M. Djamboulian. *The Implementation of Prolog*. 1993.

[7] J. B. Patrick Blackburn. *Representation and Inference for Natural Language*. 2005.