PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE

ESCUELA DE INGENIERIA - ESCUELA DE LETRAS - BIBLIOTECAS UC

DATOS ABIERTOS ENLAZADOS EN REPOSITORIO ACADÉMICO

RAFAEL CASTILLO GUERRERO

Actividad de graduación para optar al grado de
Magister en Procesamiento y Gestión de la Información

Profesor Supervisor:
PROFESOR SUPERVISOR: CLAUDIA GUTIERREZ

Santiago de Chile, Enero 2018

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE

ESCUELA DE INGENIERIA - ESCUELA DE LETRAS - BIBLIOTECAS UC

DATOS ABIERTOS ENLAZADOS EN REPOSITORIO ACADÉMICO

RAFAEL CASTILLO GUERRERO

Miembros del Comité:

PROFESOR SUPERVISOR: CLAUDIA GUTIERREZ

PROFESOR INVITADO: CÉSAR AGUILAR

DIRECTOR PROGRAMA: MAURICIO ARRIAGADA BENÍTEZ

.....

Actividad de graduación para optar al grado de

Magister en Procesamiento y Gestión de la Información

Santiago de Chile, Enero 2018

a mi Pilar fundamental, sin ella
nada de esto existiría, es mi retorno
a la vida.

AGRADECIMIENTOS

Agradezco al “Consejo Nacional de la Cultura y las Artes (CNCA), Fondo del Libro, línea Becas” quién financió mi proyecto de estudio.

Profesores: Mauricio Arriagada Director del programa, quien fue vital en las puntadas finales, a Claudia Gutiérrez por su tiempo.

También es importante agradecer a Christian Sifaqui quién fue el primero en compartir su conocimiento en LOD, Alessandro Chiaretti quién colaboró con revisiones en la parte técnica junto a Christian, además participó activamente con sus comentarios.

A SISIB de la Universidad de Chile por dar el espacio para esta aventura.

Mención especial se merece Francisco Garrido quien muy generosamente ayudo con la programación de la aplicación, lo que nos significó estar varias noches conectados vía remota para sacar adelante esta parte.

A las personas que sin proponérselo se transformaron en mentores, Javier Cancino, Isabel Maturana, Paula Muñoz, Rosa Prieto, Bruno Guerrero.

A mis amigos incondicionales esos que son escasos: Juan Alvear, Alejandro Fonseca, Rosa Leal, Seba Morán.

A mi familia, Rafael, Beatriz, Elena, Patricia, Nacho y por supuesto Verónica, quien siempre permanece en mi corazón.

Todos ellos han hecho posible que surja esta idea.

INDICE GENERAL

INDICE DE FIGURAS

INDICE DE TABLAS

RESUMEN

La investigación tiene por objetivo realizar un ejercicio semántico de datos enlazados para presentar una visualización enriquecida de datos del Repositorio de la Universidad de Chile, ejemplificando con un grupo de documentos del área forestal.

La motivación que existe para llevar adelante este estudio tiene que ver con ofrecer un sistema semántico el cual mejore la calidad y pertinencia de los metadatos que se manejan en repositorios académicos, un sistema semántico debe contar con sus metadatos altamente relacionados y estandarizados.

Para ello fue necesario realizar un estudio del estado del arte de la tecnología semántica asociada a bibliotecas, donde es posible conocer los conceptos fundamentales que abarca la disciplina de datos abiertos enlazados, así como referentes importantes de las bibliotecas en el mundo y en Chile.

Se estudiaron distintos modelos de datos aplicados al desarrollo de proyectos Linked Open Data con el fin de evaluarlos y definir la creación de un modelo propio. En este documento, se explica cómo se llegó a obtener el set de datos a procesar, junto con el proceso completo que implica trasformar un metadato en estructura de tripletas RDF. Además se describen cuáles son las herramientas que permiten desarrollar un proyecto de estas características. En este caso se optó por utilizar la herramienta de código abierto, Open Refine.

Palabras Claves: Repositorios académicos, datos abiertos enlazados, ontologías, Web semántica, marco para la descripción de recursos.

ABSTRACT

The research aims to realize a semantic exercise of linked data in order to present an enriched visualization of data from the University of Chile Repository, exemplifying with a group of documents from the forest area.

The motivation that exists to carry out this study has to do with offering a semantic system that improves the quality and relevance of metadata that are handled in academic repositories, a semantic system must have its metadata highly related and standardized.

For this purpose, it was necessary to carry out a study of the state of the art of semantic technology associated with libraries, where it is possible to know the fundamental concepts that comprise the discipline of linked open data, as well as important references of libraries in the world and in Chile.

Different data models applied to the development of Linked Open Data projects were studied in order to evaluate them and define the creation of their own model. This document explains how the dataset to be processed was obtained, along with the entire process involved in transforming a metadata into an RDF triplet structure. It also describes the tools that allow us to develop a project of these characteristics. In this case it was decided to use the open source tool, Open Refine.

Keywords: Academic Repositories, Linked Open Data, ontology, semantic Web, resource description framework.

1. INTRODUCCIÓN

1.1. Contextualización y formulación del Problema

La Universidad de Chile cuenta con 174 años de vida, fue fundada en 1842, siendo el primer rector Don Andrés Bello. De acuerdo al anuario de la universidad el énfasis ha sido desarrollar políticas de acción destinadas a resolver los problemas nacionales y regionales que afectan a Chile. Se define a sí misma como garante de la cultura clásica, humanista y secular. Es una universidad de carácter nacional y público, y trabaja en el desarrollo innovador de las ciencias y las tecnologías, las humanidades y las artes, a través de sus funciones de docencia, creación y extensión, con especial énfasis en la investigación y el postgrado. La Universidad de Chile es una de las 18 universidades del Consorcio de Universidades del Estado de Chile (CUECH[1]) y una de las 27 que conforman el Consejo de Rectores de las Universidades Chilenas (CRUCH[2]). Desde la universidad han egresado o realizado labores académicas una gran cantidad de intelectuales y destacados líderes chilenos, entre los que destacan, 20 Presidentes de la República, 179 Premios Nacionales y 2 Premios Nóbeles.

La universidad cuenta con al menos 48 bibliotecas en sus distintas facultades siendo 20 de ellas bibliotecas centrales, lo que implica que atiende a un gran número de alumnos entre pre y postgrado. La superficie total en metros cuadrados de estas unidades alcanza a los 27.536. El total de volúmenes alcanza a los 3.101.000 libros, 250 títulos de revistas impresas suscritas y 1.046 títulos de revistas electrónicas. Revistas suscritas por medio de bases de datos es de un total de 43.623, esto significa que dependiendo del título de revista es necesario pasar por alguno de los proveedores de base de datos para llegar al artículo. Las visitas Web al catálogo institucional llegan a 5.467.303. El total de artículos descargados de revistas electrónicas vía Web es de 1.911.640.

El Repositorio Académico de la Universidad de Chile nace como consecuencia de la fusión del servicio denominado Captura, donde se almacenaban artículos y trabajos de investigación por parte de los académicos, y del repositorio de Tesis de la Universidad de

[1] http://www.uestatales.cl/cue/
[2] http://www.consejoderectores.cl/

Chile, el cual almacenaba solo tesis de pre o postgrado de sus distintas facultades. Esta fusión se produjo en 2014 y como consecuencia de esta unión es que el repositorio alcanza la cifra de más de 30.000 registros almacenados con sus respectivos archivos asociados. En cuanto a tecnología, se utiliza Dspace versión 5.5 y PostgreSQL versión 9.4.1 como plataforma de almacenamiento. Actualmente no solo se ingresan las tesis de pre y post grado, sino también artículos de revista publicados por su cuerpo docente en revistas de corriente principal u otras, libros, capítulos de libros, entre otro tipo de documentos. Actualmente este sistema alcanza 43.847 registros. Se optó por conservar la definición de colecciones que existía en el anterior servicio de tesis la cual despliega las distintas facultades que posee la universidad junto con sus institutos u otros organismos de la universidad.

Dentro de las características principales de los repositorios académicos es que han permitido a las instituciones de educación superior depositar y gestionar la investigación que se lleva adelante en la universidad, como por ejemplo las tesis de los alumnos, publicaciones de los académicos u otro tipo de documentos. Quienes se encargan de administrar estas fuentes de información son las bibliotecas. De acuerdo a Duperet (2015), los repositorios son: "sistemas de información que tienen como finalidad organizar, preservar y difundir en el modo acceso abierto (Open Access) recursos científicos y académicos de las instituciones". Reafirma lo anterior (Steele y Sump-crethar, 2016) al decir que un repositorio: “es un conjunto de servicios que una universidad ofrece a los miembros de su comunidad para la gestión y difusión de materiales digitales creados por la institución y sus miembros comunitarios". Según cifras del Directory of Open Access Repositories de la Universidad de Nottingham, UK. (OpenDOAR, 2017), se muestra en el gráfico 1.1 el uso de distintos software de repositorio en Chile.

Dspace utiliza Dublin Core como sistema de metadatos, el cual fue creado en 1995 con la finalidad de poder describir recursos de información dispuestos en la Web. Es posible describir esta iniciativa como un conjunto de quince elementos o campos que puede ser usado para describir una amplia gama de recursos. Aunque en un principio se pretendía únicamente establecer el equivalente de una "tarjeta de catálogo bibliográfico". Por la rápida

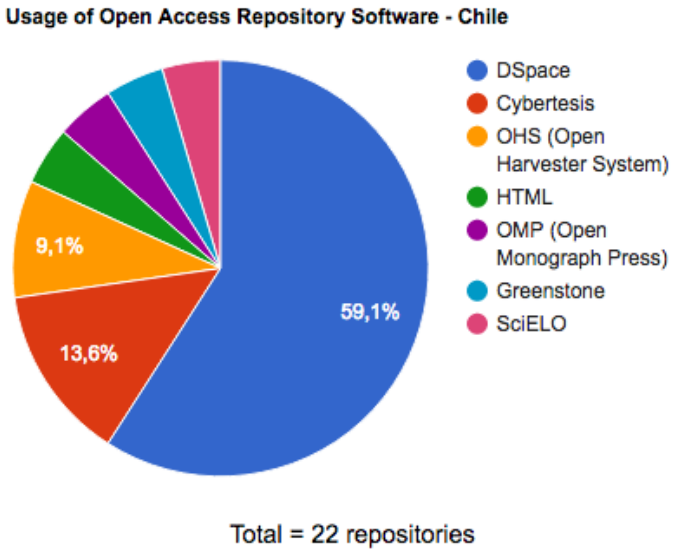


FIGURA 1.1. Repositorios según cifras de OpenDOAR y uso de software en Chile.

y fácil manera de trabajar orientada a recursos en red, el alcance de Dublin Core se expandió gradualmente durante la última década para abarcar la descripción de casi cualquier cosa. Los quince elementos básico son: Contribuidor, Cobertura, Creador, Fecha, Descripción, Formato, Identificador, Idioma, Editor, Relación, Derechos, Fuente, Objeto, Título y Tipo. Existen otros metadatos denominados Dublin Core Cualificado (ver anexo B.1), los cuales complementan a los quince mencionados en la tabla 1.1, permitiendo descripciones más específicas. Dublin Core se utiliza además como una característica fundamental de los sistemas OAI-PMH la cual permite garantizar la interoperabilidad entre sistemas de descripción de información, para lo cual utilizan Dublin Core XML como un requisito mínimo.

Si bien los repositorios académicos de acuerdo a (Gómez-Castaño, Barrueco Cruz, París-Folch, Aguilar-Lorente, y Martínez-Galindo, 2015) han permitido difundir la información patrimonial generada en las universidades bajo la premisa de acceso abierto según (Tay, 2016a): “utilizando normas como Dublin Core o permitiendo incluir metadatos personalizados, esto trae consigo algunas dificultades a la hora de buscar una normalización de contenidos o simplemente permitir que el sistema sea cosechado por otra institución".

Elemento	Definición
Contributor	Una entidad responsable por hacer una contribución del recurso
Coverage	Tópico espacial o temporal del recurso la aplicación espacial del recurso o la jurisdicción bajo la relevancia del recurso
Creator	Una entidad principalmente responsable del hacer el recurso
Date	Un punto o período de tiempo asociado a un evento en el ciclo de vida del recurso
Description	Una relación del recurso
Format	El formato de archivo medio físico o dimensiones del recurso
Identifier	Una referencia inequívoca al recurso dentro de un contexto dado
Language	Idioma del recurso
Publisher	Una entidad responsable de hacer disponible el recurso
Relation	Fuentes relacionadas
Rights	Información sobre los derechos que se tienen sobre el recurso
Source	Un recurso relacionado del cual se deriva el recurso descrito
Subject	Tema del recurso
Title	Nombre que recibe el recurso
Type	La naturaleza o género del recurso

TABLA 1.1. 15 elementos básicos de Dublin Core

Sistemas como Dspace han utilizado una combinación de tecnologías que le permiten al usuario final realizar búsquedas de información obteniendo resultados que se complementan a través de un sistema de descubrimiento el cual utiliza Apache Solr[3] como base. Este motor de búsqueda hace entrega de listado de facetas con los distintos campos definidos como parte del descubridor entregando información relacionada a la búsqueda efectuada, entre otras tareas. Es decir que si buscamos por *eucalyptus globulus*, los campos facetados nos mostrarán autores que han escrito sobre el tema, complementado con las fechas de dichos textos y además de los distintos tipos de material que cubren el tópico. Sin embargo si la información no se ha ingresado o se ha omitido no será parte del ecosistema de información resultante para la búsqueda. Tal como lo expresa (Bui y Park, 2006) “la calidad de los registros de metadatos en una biblioteca digital tiene un efecto crítico en el acceso y recuperación de la información".

[3] `https://es.wikipedia.org/wiki/Apache_Solr`

Los repositorios académicos se enfrentan a una serie de dificultades como lo son: competir directamente con los buscadores de la Web en lo que a búsqueda de información se refiere, tal y como lo plantea (Rodríguez-Bravo, Simoes, Vieira-de Freitas, y Frías, 2017): “Diversas investigaciones muestran que cada vez es más frecuente que los usuarios no comiencen el descubrimiento de información en el portal de la biblioteca sino en Google, Google Scholar o plataformas similares".

Además de lo anterior los repositorios según menciona (Rodríguez-Bravo et al., 2017) “no parecen tampoco tener interés para los jóvenes investigadores entrevistados para el proyecto Harbingers. Los resultados muestran que la mayoría de ellos no tiene hábito de autoarchivar. Quienes tienen constancia de la existencia de un repositorio en su institución, suelen responder que son los bibliotecarios quienes se encargan de poner en acceso abierto sus documentos”. Es el caso de la Universidad de Chile donde personal bibliotecario y especializado son los que ingresan información al sistema de repositorio académico.

Por otra parte algunas dificultades importantes de los repositorios tal y como plantea (Tay, 2016c): “carecen de una alarmante falta de normalización de datos, y que esa laxitud –uso del mínimo de etiquetas de Dublin Core- les está perjudicando porque la recuperación resulta en un conjunto inconsistente de ítems". Profundizando la idea anterior es importante destacar que Dublin Core no se presenta como la mejor opción al momento de describir metaetiquetas de artículos académicos por parte de Google Scholar siendo Highwire Press la más aceptada de acuerdo a su “Inclusion Guidelines for Webmasters"(Scholar, 2017).

Por otro lado, según (Breitbach, 2016) apunta como debilidad la falta de consistencia en los metadatos y la ausencia de contexto en la relación de éstos con las disciplinas. Señala que: “los servicios de descubrimiento agregan datos con el fin de normalizarlos en un único índice y que ello trae como consecuencia que el valor de los metadatos disminuya. Estos sistemas no pueden competir con las bases de datos disciplinares en pertinencia de los resultados. La misma situación se ha puesto de relieve para los repositorios institucionales".

La tecnología semántica se observa como una alternativa de explorar ya que permitiría la estructuración de datos de manera relacionada en su contenido, tal y como lo expresa

(Southwick, 2015) quien plantea tres motivos por los cuales utilizar LOD: “en primer lugar, mediante la adopción de los conceptos y tecnologías LOD, podríamos romper los silos en los que residen nuestros metadatos de colecciones digitales; en segundo lugar, interconectar nuestros datos con los datos relevantes de otros proveedores de datos. Por último, la interconexión de los datos permitiría a los usuarios realizar búsquedas continuas de la información pertinente, independientemente de dónde se haya generado o quién la haya generado”.

1.2. Motivación

El presente ejercicio busca poner en discusión a los sistemas de metadatos utilizados para describir información en repositorios institucionales, pues tal y como lo expresa (Bui y Park, 2006), “la calidad de los registros de metadatos en una biblioteca digital tiene un efecto crítico en el acceso y recuperación de la información".

De acuerdo a (Universidad de Chile, 2017) a través de su sistema de recolección de datos Repositorio Latinoamericano: en América Latina se utiliza Dublin Core principalmente como sistema de metadatos en repositorios institucionales incluidos los académicos. Sin embargo, estos sistemas no incorporan metadatos asociados al contenido que se intenta describir -más allá, claro está, del campo de materias o palabras claves. Más bien, estos sistemas se encuentran centrados en el formato, ya sea para la descripción de libros, así como para la de revistas, artículos de publicaciones periódicas, tesis u otro tipo de documentos como puede ser el audiovisual. Esta falta de descripción de contenidos se aprecia en las tesis de alumnos de pregrado y postgrado, las cuales en un repositorio académico pueden cubrir temáticas muy diversas como Ciencias Médicas o Geología. Por ello se pretende buscar una solución a través de un sistema que no sólo permita agregar más metadatos, sino que posibilite establecer relaciones con un mayor nivel de semántica, ya que la Web semántica nos ofrece otro nivel aún más enriquecido tal y como lo plantea (Konstantinou, Houssos, y Manta, 2014): “la respuesta a las consultas se puede realizar utilizando sistemas convencionales, hasta el punto de que las consultas estándar basadas en palabras claves se evalúan de forma eficiente y los resultados respectivos pueden abarcar varios repositorios,

ofreciendo una solución básica. Por el contrario, la consulta de grafos mediante patrones de grafo permite realizar consultas mucho más complejas, lo que permite la formación de consultas más descriptivas". Por lo tanto, es posible alcanzar niveles de respuesta que un sistema convencional no ofrece.

Tal como plantea (Chen, 2017) “basados en los principios de Linked Open Data, los registros heredados pueden ser deconstruidos en datos LOD que pueden ser enriquecidos con información variada agregando con otros recursos externos y sus contextos en la Web". Es importante considerar también las ventajas de trabajar en un espacio interconectado y enriquecido como lo explica (Di Noia et al., 2016): “vincular estos metadatos con los conjuntos de datos en la nube Linked Data (GeoNames, DBpedia, FOAF, etc.) mejora enormemente la reutilización e integración de diversas bibliotecas digitales. La eficacia en el uso de los conjuntos de datos LOD para anotar recursos digitales en una biblioteca digital también se ve reflejada en algunos casos de éxito, como el de la Biblioteca Nacional Alemana o la Biblioteca Nacional Británica, así como en el caso de la Biblioteca Nacional de España y la Biblioteca Nacional de Francia". Es importante destacar los beneficios de optar por sistemas de datos abiertos enlazados, pues según (Di Noia et al., 2016):“hay muchos beneficios en el uso de metadatos vinculados para bibliotecas digitales, entre ellos: apertura e intercambio de metadatos, facilidad en el descubrimiento de información, identificación de patrones de uso de recursos, navegación basada en facetas y metadatos enriquecidos con enlaces".

Es importante destacar lo que plantea (Goddard y Byrne, 2010): “la Web semántica se basa en metadatos altamente estructurados que permiten a los ordenadores comprender las relaciones entre objetos. Los estándares Web semánticos son complejos y difíciles de conceptualizar, pero ofrecen soluciones a muchos de los problemas que afectan a las bibliotecas, incluyendo la búsqueda Web precisa, control de autoridad, clasificación, portabilidad de datos y desambiguación". Además las autoras aseguran que: “los datos vinculados también podrían ayudarnos a descubrir ese santo grial de las tecnologías bibliotecarias: la búsqueda federada inteligente. Si todos nuestros proveedores de recursos electrónicos expusieran sus datos en RDF y utilizaran *hubs* de enlace para aplicar un vocabulario comúnmente

controlado, entonces podríamos eliminar muchos de los problemas asociados actualmente con los buscadores federados como la falta de granularidad, la incapacidad para soportar consultas sofisticadas, el pobre ranking de relevancia, el desdoblamiento inexacto y la lentitud". Esto pues los buscadores tradicionales se basan principalmente en búsquedas por cadenas de caracteres a través de consultas SQL y no por la relación semántica que existe entre los distintos individuos que componen un sistema de datos abiertos enlazados.

De acuerdo a (Southwick, 2015) tres beneficios de LOD son: “en primer lugar, mediante la adopción de los conceptos y tecnologías LOD, podríamos romper los silos en los que residen nuestros metadatos de colecciones digitales y, en segundo lugar, interconectar nuestros datos con los datos relevantes de otros proveedores de datos. Por último, la interconexión de los datos permitiría a los usuarios realizar búsquedas continuas de la información pertinente, independientemente de dónde se haya generado o quién la haya generado".

La tecnología semántica amplía las capacidades de un repositorio tradicional (ver punto 5.7) , ya que permite visibilizar información que subyace al momento de realizar una búsqueda, esto se logra a través de lo que se denomina inferencia semántica, la cual permite relacionar contenidos a través de la estructura sujeto-predicado-objeto. Por ejemplo, al obtener un registro del repositorio que en su contenido trata sobre la VIII Región, se podrá saber y relacionar otros temas del área forestal que estén presentes en otros registros que contengan la misma región. Un repositorio tradicional no entrega esta información, sería necesario formular una consulta SQL, la cual sería necesariamente compleja de realizar por un usuario.

Por otro lado el gran aporte de la tecnología semántica es que permite vincular información desde fuentes de datos externas para incorporarla en los registros de un repositorio, es así como en este ejercicio es posible visualizar en un registro del repositorio académico, la definición para nombres científicos de las especies forestales, las que son leídas desde DBpedia. También es posible obtener los datos que complementan la información de una región, como es la superficie, número de habitantes, coordenadas geográficas entre otras.

A su vez la información que posee el repositorio académico al estar disponible como datos abiertos enlazados, podrían ser integrados por otras fuentes de datos, es decir nutrir a otras instituciones con información propia de la universidad y además fuente autorizada.

Por lo expuesto anteriormente, es importante que las universidades no sólo sean consumidoras de datos, sino que también puedan proveer de datos proporcionando acceso a su acervo cultural depositado en tesis. En esta línea se propone desarrollar un sistema que permita gestionar la información patrimonial de una universidad utilizando un camino distinto a los tradicionales repositorios académicos, para lo cual se busca elevar el nivel de exhaustividad en la descripción de información y su relación, a través de sistemas de tecnología semántica. Los datos estructurados semánticamente ayudan a los motores de búsqueda a analizar y comprender el contenido web, entregando con esto precisión en la recuperación de información hacia los usuarios, además de generar nuevos puntos de entrada a las tesis de la universidad.

1.3. Herramientas

Las herramientas utilizadas en el presente proyecto para lograr los objetivos descritos en el punto anterior son las siguientes:

- Extracción de información a través de un cosechador de registros en formato Dublin Core a través del software Marcedit, en versión para MacOS. Este programa permitió extraer el set de datos desde el repositorio académico de la Universidad de Chile.
- Limpieza, extracción y generación de nuevos set de datos a través de OpenRefine. Como lo sugiere (Van Hooland y Verborgh, 2014): “se incorporó el complemento para OpenRefine denominado rdf-extension el cual nos permite definir la estructura de los archivos RDF ya que se incorpora un sistema de reconciliación para establecer vínculos con otras fuentes de datos".
- Revisión de información en cualquier momento sin la limitación de tiempo y lugar a través de hojas cálculo de Google Drive.
- Lenguaje de programación Php para generar el sitio de prueba del sistema.
- Open Link Virtuoso versión *open source*, como sistema de *Triple Store* para base de datos de RDF.

1.4. Alcances y limitaciones

El siguiente proyecto busca realizar una investigación teórica y práctica que permita explorar los pasos a seguir para obtener un archivo RDF base para la creación de un sistema de datos abiertos enlazados que potencialmente se integre al Repositorio Académico de la Universidad de Chile, tomando como punto de partida un set de datos recolectados desde el repositorio a través del software MarcEdit versión MacOs.

Se continua con una revisión de la literatura que permita comprender las ventajas y desventajas de los repositorios institucionales, así como verificar cuales son las principales

dificultades que posee dicho sistema. Por otro lado la revisión bibliográfica abarcó información sobre datos abiertos enlazados, distintas experiencias al respecto, implementaciones, entre otras.

Posteriormente se cargaron los datos en OpenRefine con el fin de realizar análisis y limpieza de la carga. Luego se procedió a utilizar Google Drive para incorporar nuevos metadatos así como para iterar sobre el análisis efectuado, esta herramienta tiene la ventaja de permitir el trabajo remoto a diferencia de otras.

Para continuar con el trabajo se revisan una serie de ontologías, donde se seleccionan las más pertinentes. Una vez concluido lo anterior es posible comenzar a modelar en distintas iteraciones lo que será la trasformación de metadatos en *Linked Open Data* es decir archivos RDF.

Se procede a la habilitación de *Open Link Virtuoso*, donde se realiza la carga de los distintos archivos RDF.

Finalmente se programa con la colaboración de un experto programador lo que será la interfaz de consulta del sistema.

A continuación se presenta una tabla resumen 1.2 de las etapas mencionadas:

Etapa	Definición	Tiempo	Tecnología
Planificación	Definición de set de datos		
	Recolección del set de datos	1 día	MarcEdit
	Revisión de literatura	todo el proyecto	
Procesamiento	Carga de datos en herramienta	1 día	OpenRefine
	Análisis de datos	2 semanas	
	Limpieza de datos	4 semanas	OpenRefine
	Incorporación de metadatos	4 semanas	OpenRefine
	Revisión de información	3 semanas	Google Drive
	Nueva limpieza de datos	2 semenas	OpenRefine
Generación de RDF	Definición de ontologías	3 semanas	
	Tranformación de metadatos a LOD	2 días	OpenRefine y RDF extension
	Revisión de resultados		SublimeText y Firefox
Publicación en servidor	Carga de archivos RDF	1 día	OpenLink Virtuoso
Interfaz	Diseño y desarrollo de interfaz de consulta	4 semanas	Php y OpenLink Virtuoso

TABLA 1.2. Resumen de las etapas del proyecto

Se utilizarán los conceptos Linked Open Data, LOD o datos abiertos enlazados indistintamente como sinónimos.

1.5. Objetivos

1.5.1. Objetivo General

El objetivo general de este ejercicio es proveer un modelo de metadatos que incorpore sistemas de datos abiertos enlazados, a través de tecnología semántica al Repositorio Académico de la Universidad de Chile. Para poder realizar este ejercicio se trabajará con una muestra de datos del área forestal que permite incorporar metadatos de geolocalización a las búsquedas.

1.5.2. Objetivos Específicos

- Definir características mínimas para la elección del set de datos a trabajar.
- Identificar e incorporar campos que permitan potenciar la información que se obtiene del repositorio Académico.
- Enriquecer los datos bibliográficos a través de un sistema de datos abiertos enlazados que incorpore relaciones semánticas con otras fuentes de datos.
- Proponer una nueva visualización de datos que permita buscar y ver las relaciones de información en el sistema y proporcionar un mejor acceso al conocimiento de tesis y publicaciones de la universidad.

1.6. Estructura del proyecto

En los siguientes capítulos se abordará una serie de temáticas que dan sustento al proyecto: en el **capítulo dos** se revisa cuál es el estado del arte de los *datos abiertos enlazados* en el mundo y en Chile en cuanto a bibliotecas se refiere, luego en el **capítulo tres** se revisará cual fue la metodología empleada para realizar la conversión de la muestra del set de datos del repositorio académico de la Universidad de Chile desde Dublin Core XML a RDF. En el **capítulo cuatro** se explica cómo se realizó una aplicación a un caso de estudio, para continuar con el **capítulo cinco** es posible observar cuáles son los resultados del proyecto junto con un breve análisis de lo obtenido. Finalmente se presentan las **conclusiones** a las que se han llegado después de todo el ejercicio.

2. ESTADO DEL ARTE

2.1. Datos abiertos enlazados

Best Practices for Publishing Linked Data (W3C, 2014a) define a datos abiertos enlazados como: “conjunto de mejores prácticas para publicar e interconectar datos estructurados para el acceso tanto por humanos como por máquinas".

De acuerdo a (Heath y Bizer, 2011) el término Linked Data se refiere a un conjunto de mejores prácticas para publicar e interconectar datos estructurados en la Web. Estas mejores prácticas fueron introducidas por Tim Berners-Lee en su nota de arquitectura Web Linked Data y se han dado a conocer como principios de Linked Data.

Por otro lado (Southwick, 2015) menciona cuatro conceptos fundamentales para entender los datos abiertos enlazados:

1. Tripletas (también llamadas declaraciones).
2. Identificadores uniformes de recursos (URI).
3. Marco de Descripción de Recursos (RDF).
4. Sparql, un lenguaje para acceder y gestionar LOD.

Tripletas o declaraciones En un ambiente LOD los datos se expresan en declaraciones llamadas tripletas, poseen tres componentes: sujeto - predicado - objeto.

Ejemplo de declaraciones:

Araucaria araucana se ubica en el Parque Nacional Conguillío

Araucaria araucana = *sujeto*

se ubica en = *predicado*

Parque Nacional Conguillío = *objeto*

Identificadores únicos de recurso Los identificadores únicos de recurso permitirán la vinculación entre los datos dentro y a través de los conjuntos de datos. Por esta razón, es esencial reutilizar los URI existentes siempre que sea posible.

Por otro lado (Heath y Bizer, 2011) explica que: “el protocolo HTTP es el mecanismo de acceso universal de la Web. En la Web clásica, las URIs HTTP se utilizan para combinar la identificación única global con un mecanismo de recuperación simple y bien entendido". Así, el segundo principio Linked Data aboga por el uso de URI HTTP para identificar objetos y conceptos abstractos, permitiendo que estos URIs sean referenciados (es decir, consultados) sobre el protocolo HTTP en una descripción del objeto o concepto identificado ver ejemplo de figura 2.1 .

Ejemplo de URI:

`http://id.loc.gov/authorities/names/no2016122675`

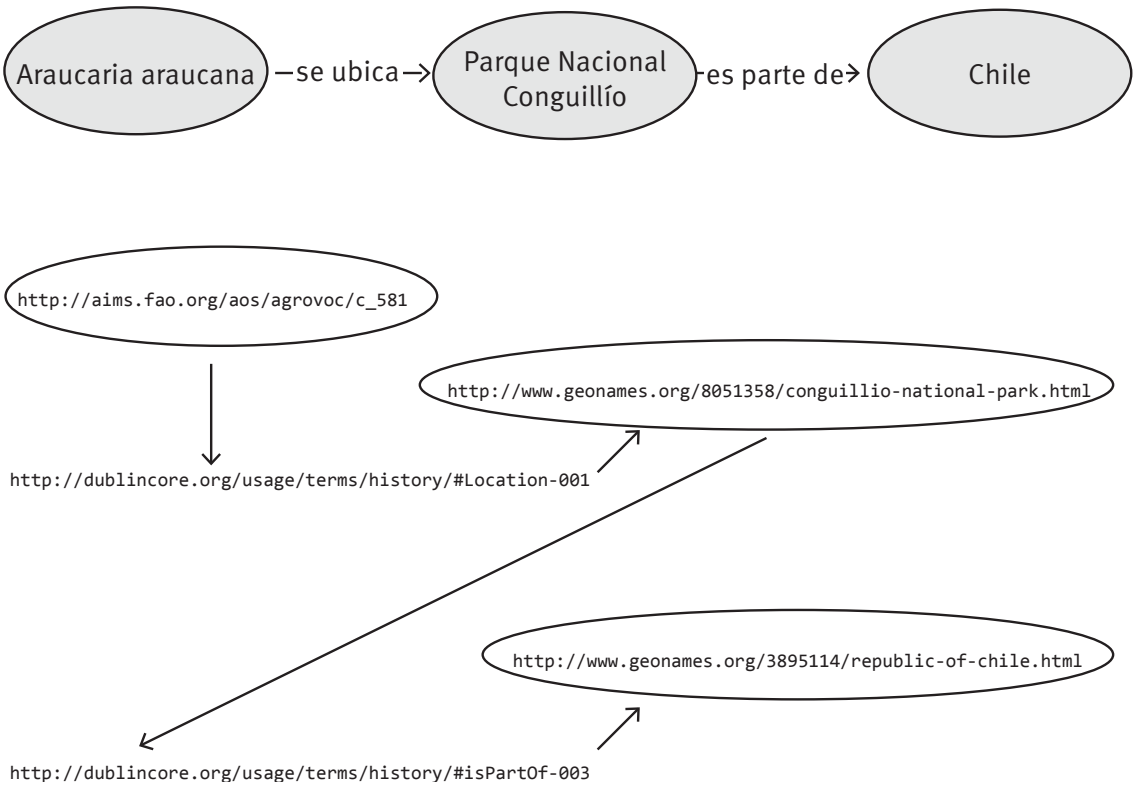


FIGURA 2.1. Ejemplo de representación de tripletas legible por humano y por máquina.

Resource Description Framework

El consorcio de la Web (W3C, 2014b) define Resource Description Framework (RDF) como: “un lenguaje para representar información sobre recursos en la World Wide Web. Está especialmente destinado a representar metadatos sobre recursos Web, como el título, autor y fecha de modificación de una página Web, información sobre derechos de autor y licencias sobre un documento Web, o el calendario de disponibilidad de algún recurso compartido. El RDF está pensado en que la información necesita ser procesada por aplicaciones o máquinas, en lugar de ser mostrada sólo a las personas. RDF proporciona un marco común para expresar esta información de manera que pueda intercambiarse entre aplicaciones sin pérdida de significado. La capacidad de intercambiar información entre diferentes aplicaciones significa que la información puede ponerse a disposición de otras aplicaciones distintas de aquellas para las que se creó originalmente".

Lenguaje Sparql: De acuerdo a (Heath y Bizer, 2011) sparql es un acrónimo utilizado para sparql Protocol y RDF Query Language. Una característica importante de sparql es que puede consultar datos en cualquiera de los formatos de serialización de RDF, así como en cualquier combinación de ellos. El aspecto de protocolo de sparql implica que las máquinas pueden intercambiar consultas y resultados. Estas dos características hacen de sparql un lenguaje muy poderoso. Por otro lado (Wood, Zaidman, Ruth, y Hausenblas, 2014) plantea que cada base de datos necesita un lenguaje de consulta. sparql es a los datos RDF como SQL es a una base de datos relacional. sparql es el lenguaje de consulta para datos estructurados en la Web, específicamente los datos accesibles en formatos RDF o representables como tales. Sparql es por lo tanto el lenguaje de consulta para Linked Data. El propósito principal de sparql es proporcionar un lenguaje formal en el cual las preguntas significativas pueden ser formuladas.

2.2. Situación actual en datos abiertos enlazados en bibliotecas del mundo

Existen distintas iniciativas que han implementado Linked Open Data en el mundo, entre las que destaca la iniciativa de la Biblioteca Nacional de Francia, España, Inglaterra, Alemania, Europeana, Biblioteca del Congreso en Estados Unidos, entre otras.

Biblioteca Nacional de Francia[1]**:** El proyecto data.bnf.fr se esfuerza para que los datos producidos por la Biblioteca Nacional de Francia sean más útiles en la Web. Reúne varios recursos de la BnF y recursos externos en páginas dedicadas a un autor, una obra o un tema. Estas páginas organizan los contenidos, vínculos y servicios Web que ofrece BnF. Disponible en línea desde julio de 2011, data.bnf.fr sigue evolucionando y ampliándose.

El objetivo es presentar las colecciones de la BnF y proporcionar un centro entre diferentes recursos. http://Data.bnf.fr está diseñado para soportar las otras aplicaciones de BnF. El proyecto pertenece a la política de la BnF de convertirse en parte de la Web de datos y adoptar los estándares de la Web Semántica. Cuenta con un endpoint en http://data.bnf.fr/sparql/.

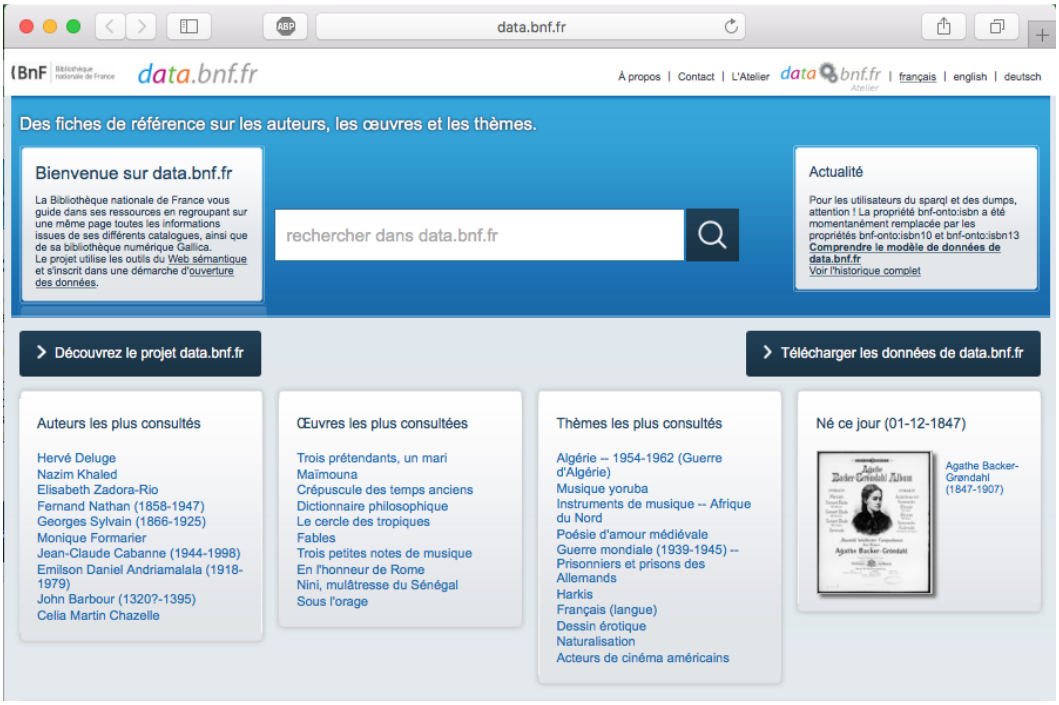


FIGURA 2.2. Biblioteca Nacional de Francia.

[1] http://data.bnf.fr/about

Biblioteca Nacional de España[2] **:** Ha sido desarrollado por la Biblioteca Nacional de España y el Ontology Engineering Group para el público objetivo de usuarios de bibliotecas y desarrolladores de TI especializados en tecnologías de Web Semántica. Más bien desde un punto de vista experimental.

De acuerdo a lo que aparece en el sitio Web se menciona lo siguiente: "se trata de un proyecto experimental para usuarios finales e investigadores, que tiene como objetivo hacer los datos bibliográficos más accesibles y más fáciles de explorar de una manera totalmente diferente de los catálogos tradicionales, ofreciendo una experiencia de navegación completamente nueva que es completamente diferente del enfoque tradicional, reuniendo los diversos recursos de la biblioteca y enriquecer sus propios datos con información externa".

FIGURA 2.3. Biblioteca Nacional de España.

[2] http://datos.bne.es/inicio.html

Biblioteca Nacional Británica[3]**:** ha desarrollado una versión de la Biblioteca Nacional Británica que está disponible como Linked Open Data a través de una plataforma de TSO. El Linked Open BNB incluye libros publicados y futuros, así como publicaciones periódicas.

A diferencia de otros proyectos la Biblioteca Nacional Británica ha publicado los modelos de ontología que utilizó para la estructuración de libros y publicaciones periódicas, todo esto actualizado a 2017. Es posible acceder al endpoint a través de `http://bnb.data.bl.uk/sparql`.

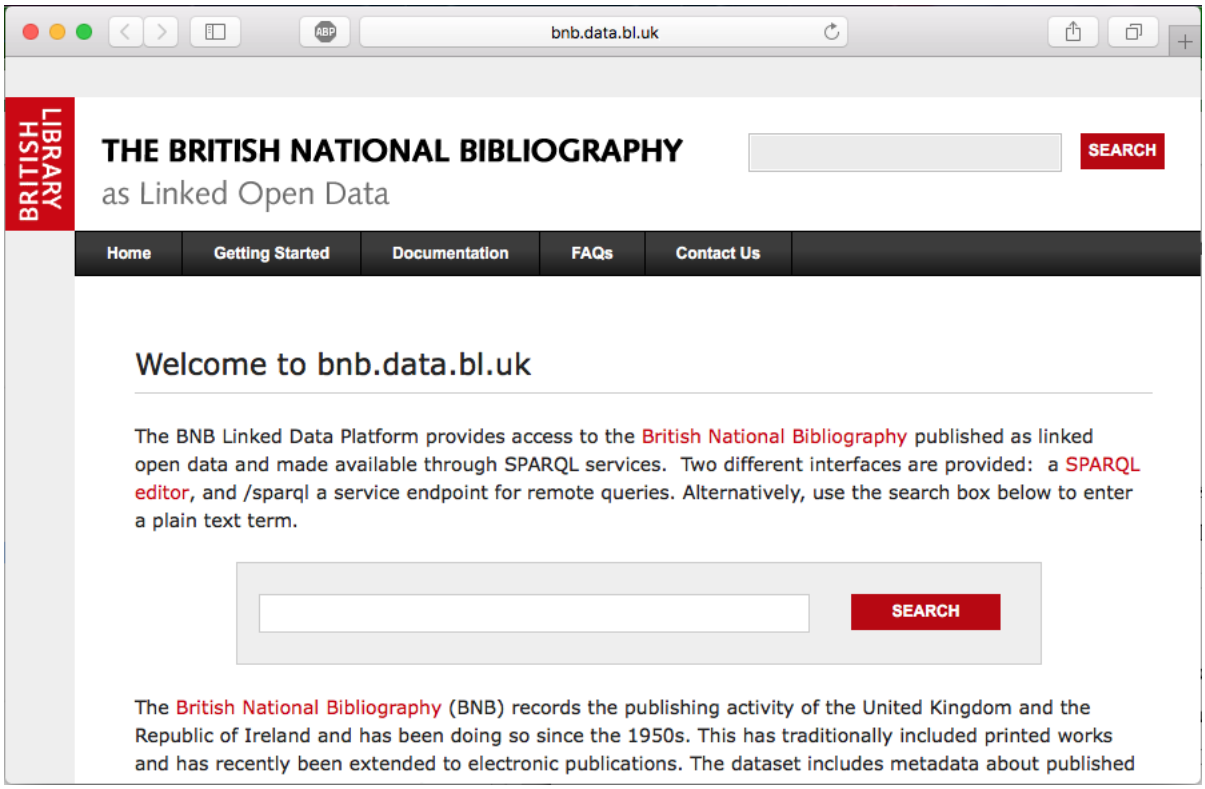


FIGURA 2.4. Biblioteca Nacional Británica.

[3]`http://www.bl.uk/bibliographic/datafree.html`

Biblioteca Nacional de Alemania[4] crea un servicio de datos vinculados, que permite el uso de toda la información bibliográfica nacional incluyendo todos los datos estándar de la comunidad de la Web Semántica en el largo plazo. Se quiere hacer que este servicio de información contribuya a la infraestructura de la información en el mundo y por lo tanto llegar a ofrecer un sistema que sea requisito previo como modelo para los modernos servicios web comerciales y no comerciales.

Desde 2010, la Biblioteca Nacional Alemana coloca sus datos a disposición como RDF a través del Servicio de Linked Data. El servicio se desarrolla de forma continua y optimizado para el nivel técnico, sustantivo y de organización. Además del desarrollo de su propio servicio, la Biblioteca Nacional de Alemania participa activamente en la iniciativa de transición hacia BIBFRAME Alemania. En este sistema no es posible realizar consultas vía endpoint, sin embargo disponen sus set de metadatos para descargar desde: `http://datendienst.dnb.de/cgi-bin/mabit.pl?userID=opendata&pass=opendata&cmd=login`

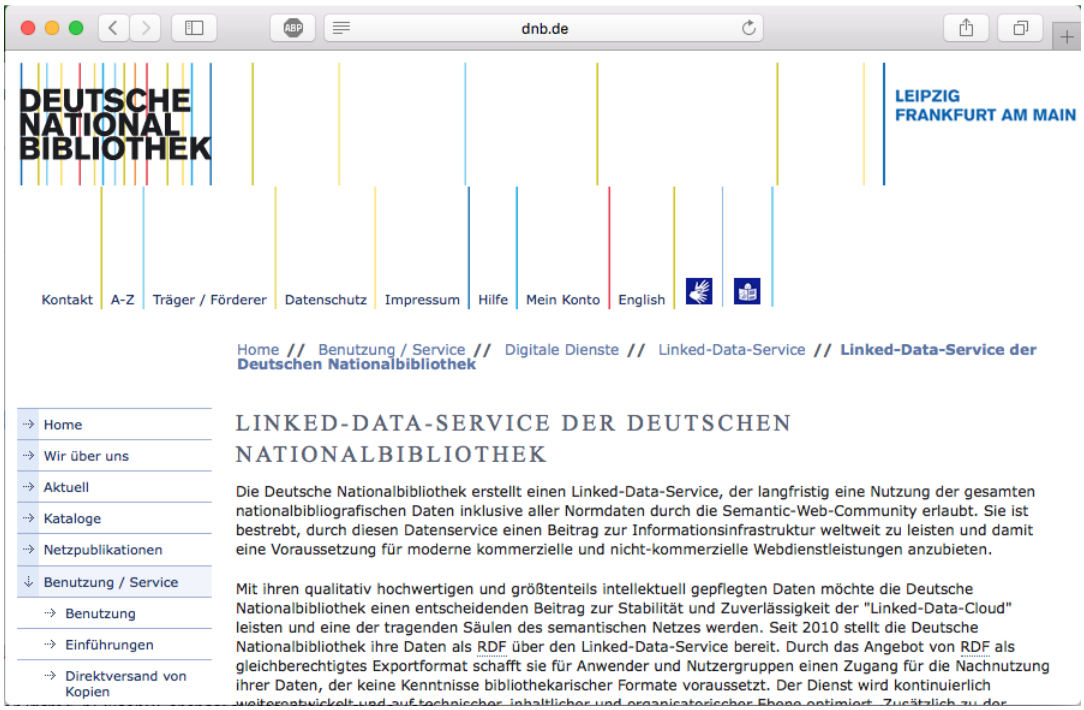


FIGURA 2.5. Biblioteca Nacional de Alemania.

[4] `http://www.dnb.de/DE/Service/DigitaleDienste/LinkedData/linkeddata_node.html`

Biblioteca Nacional de Suecia[5]**:** en 2011 la Biblioteca Nacional Sueca y los archivos de autoridad fueron puestos en libertad en el dominio público. La bibliografía nacional y los datos de autoridad es parte de Libris, el Catálogo Sueco de la Unión, y el objetivo a largo plazo es liberar toda la base de datos bajo una licencia abierta.

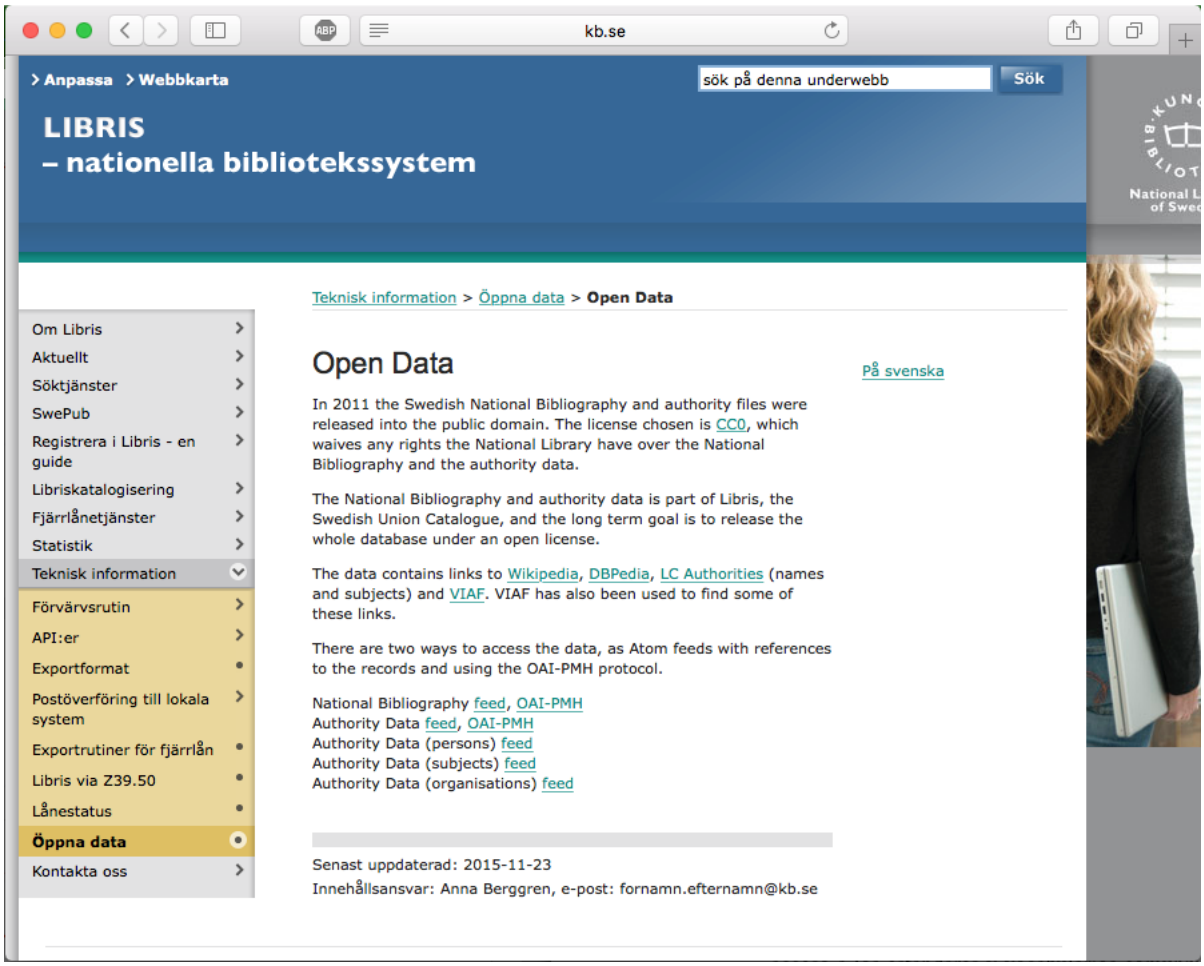


FIGURA 2.6. Biblioteca Nacional de Suecia.

[5]http://www.kb.se/libris/teknisk-information/Oppen-data/Open-Data/

Europeana[6]**:** es la biblioteca digital europea, comenzó como piloto experimental en febrero de 2012 con un pequeño número de proveedores de datos que se comprometieron en una fase temprana con la iniciativa de promover datos abiertos. En octubre de 2012, un gran subconjunto del conjunto de datos de Europeana (metadatos sobre 20 millones de textos, imágenes, videos y sonidos) se transformó en datos enlazados y se puso a disposición. Más recientemente, dentro del proyecto Creative Europeana, se desarrolló un nuevo piloto, que fue organizado por Ontotext Corp., que contenía todo el conjunto de datos de Europeana en ese momento (metadatos en unos 36 millones de registros).

Todos los conjuntos de datos Europeana pueden ser explorados y consultados a través de un sparql endpoint.

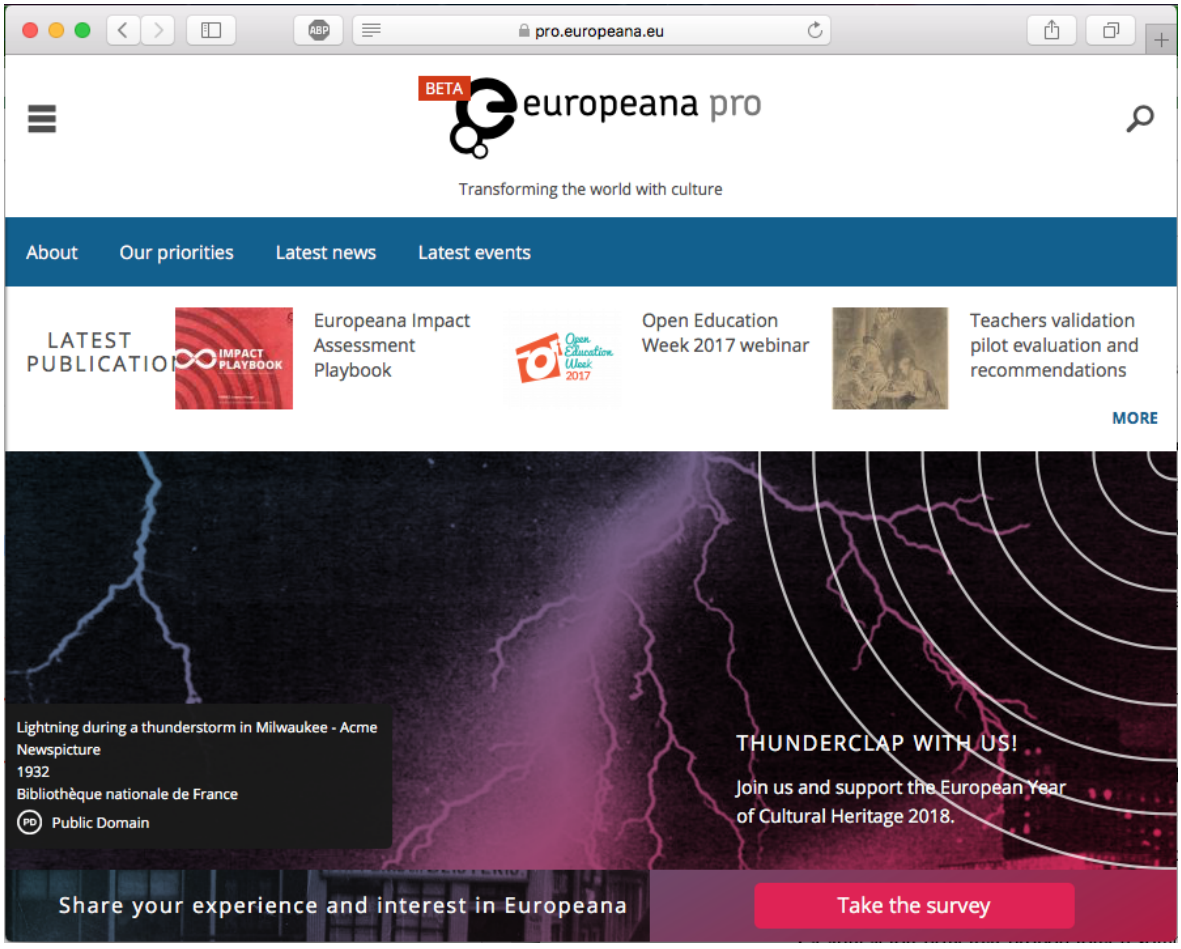


FIGURA 2.7. Europeana.

[6]http://labs.europeana.eu/api/linked-open-data-introduction

Biblioteca del Congreso de Estados Unidos[7]**:** el Servicio de Datos Enlazados proporciona acceso a los estándares y vocabularios comúnmente promulgados por la Biblioteca del Congreso. Esto incluye los valores de los datos y los vocabularios controlados que el sistema almacena y dispone a la comunidad.

La aplicación principal proporciona resolución y entrega acceso a datos y vocabularios asignando URIs. Cada vocabulario posee un URI, al igual que cada valor de datos dentro de él.

Los URIs accesibles en id.loc.gov solo enlazan con datos de autoridad, es decir, vocabularios controlados y los valores dentro de ellos. Por lo tanto, los usuarios no encontrarán identificadores para recursos bibliográficos electrónicos.

Por otra parte la Biblioteca del Congreso de Estados Unidos junto a Zepheira una empresa dedicada a los proyectos de Linked Open Data ha desarrollado el estándar que reemplazará al sistema de metadatos Marc, denominado BIBFRAME[8]. Es posible descargar masivamente una serie de datos desde: `http://id.loc.gov/download/` pero no realizar consultas vía endpoint.

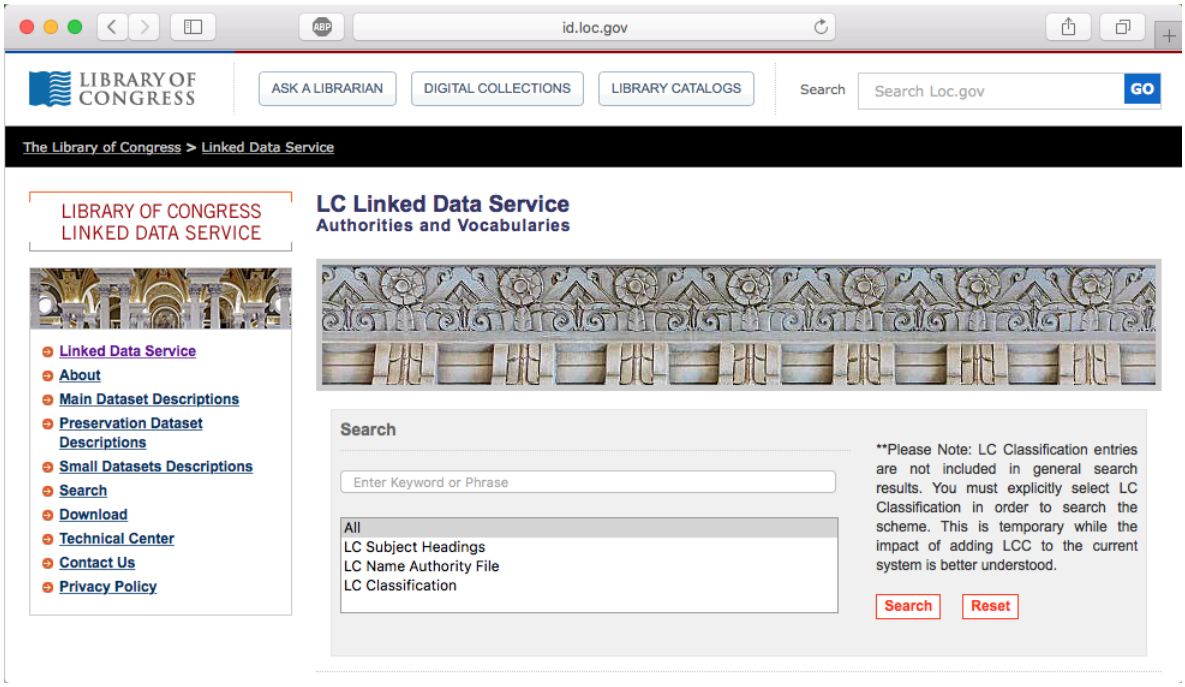


FIGURA 2.8. Biblioteca del Congreso de Estados Unidos.

[7] `http://id.loc.gov/about/`
[8] `https://www.loc.gov/bibframe/`

2.3. Datos abiertos enlazados en Chile

La experiencia de datos abiertos vinculados en Chile, prácticamente se reduce a dos que han alcanzado el tramo de cinco estrellas, Biblioteca del Congreso Nacional de Chile, y la Universidad de Chile.

FIGURA 2.9. Datos Universidad de Chile.

Biblioteca del Congreso de Chile[9]**:** fue la primera experiencia en Chile sobre el tema, inició el proyecto en 2011 y continuó desarrollando hasta 2013, a partir de ese momento se implementaron al menos tres iniciativas Historia de la Ley, Labor Parlamentaria y LeyChile. Este proyecto creó su propia ontología para definir la estructura de información acerca de parlamentarios chilenos, además de incorporar otras ontologías que apoyan la vinculación con fuentes externas. Es posible consultar sus datos a través del endpoint http://datos.bcn.cl/sparql.

[9] `http://datos.bcn.cl/es`

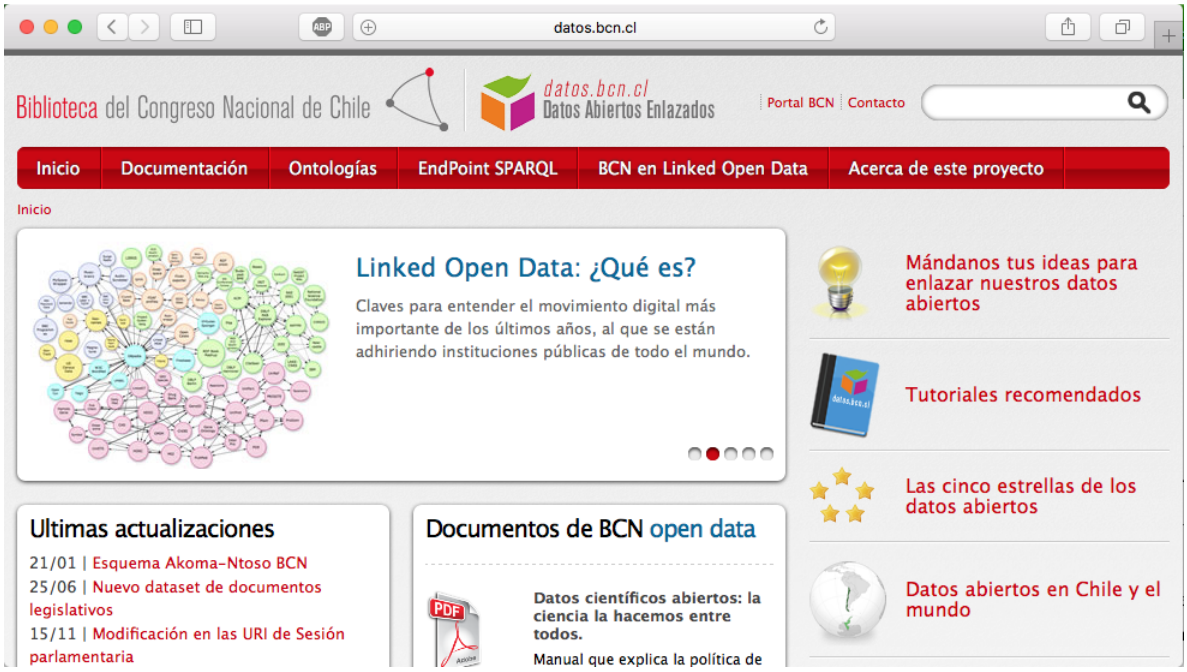


FIGURA 2.10. Biblioteca del Congreso de Chile.

2.4. Dspace como repositorio de datos abiertos enlazados

Duraspace, a través de su software Dspace, de acuerdo a (Konstantinou, Spanos, Houssos, y Mitrou, 2014) utiliza PostgreSQL[10] en el backend como base de datos para almacenar sus metadatos. Este programa introduce la opción de migrar el repositorio alojado en su sistema a Linked Open Data en mayo de 2015 en su versión 5.x. Lo mantiene en su última versión oficial 6.x de septiembre de 2016, a través de la cual podemos realizar consultas a un Sparql endpoint. No obstante es necesario realizar una serie de configuraciones además de instalar otros programas computacionales como Apache Fuseki -un servidor de tripletas- tal como lo sugiere la documentación para la instalación de Dspace. La filosofía que está detrás de esto lo expresa (Duraspace, 2016) de la siguiente forma: “para la mayoría de los repositorios, al menos para los repositorios de Open Access, es muy importante compartir su contenido almacenado".

¿Qué dificultades plantea la utilización del sistema que nos propone Dspace para LOD? Al utilizar Dspace para generar archivos RDF directamente tal y como queda la instalación del sistema es necesario considerar lo siguiente que nos plantea (Konstantinou, Spanos, et al., 2014): “en primer lugar, DSpace Application Profile amplía el conjunto de elementos de metadatos Dublin

[10] `https://www.postgresql.org/`

Core original con varios calificativos personalizados que no tienen una contraparte exacta en el vocabulario Dublin Core".

Otro punto a considerar según (Konstantinou, Spanos, et al., 2014): “es el hecho de que Dublin Core es un vocabulario de propósito general para describir los recursos web y, como resultado, la semántica exacta de sus elementos se deja deliberadamente subespecificada".

Si bien Dspace podría mapear sus metadatos con alguna ontología distinta al esquema de metadatos Dublin Core, el mismo (Konstantinou, Spanos, et al., 2014), expresa: “se podrían realizar mapeos para otro tipo de material (por ejemplo, contenidos audiovisuales, exposiciones en museos, material didáctico)". Incluso (Konstantinou, Spanos, et al., 2014) nos dice que:“puede elegir propiedades más especializadas de otros vocabularios que representen coincidencias conceptuales más cercanas que los elementos dc".

Es importante destacar que Dspace podría ser suficiente para crear un sólo archivo RDF con todo un repositorio contenido en él, pero por otro lado no es capaz de entregar set de datos disgregados o separados de acuerdo a sus características, como pueden ser los autores, o contribuidores o incluso materias, lo cual es justamente lo que intenta comprobar este proyecto, buscar una respuesta a la pregunta ¿qué comportamiento tiene un repositorio en sus opciones de búsqueda al contar con información normalizada y relacionada en sus distintos grafos RDF?.

Finalmente y como plantea (Cole, Han, Weathers, y Joyner, 2013): “el núcleo simple de Dublin Core, con solo 15 elementos y sin posibilidad real de vincular las URI a los valores de las cadenas, no se considera generalmente suficientemente expresivo (por ejemplo, no hay manera de distinguir entre diferentes clases de identificadores, temas, nombres corporativos y personales, etc.)". Se conoce que Dspace utiliza como sistema de metadatos base a Dublin Core.

3. METODOLOGÍA

Se ha optado por el estudio de registros bibliográficos pertenecientes al Repositorio Académico de la Universidad de Chile. Se seleccionó la colección de tesis perteneciente a la Facultad de Ciencias Forestales, de ella se extrajo una muestra de registros a utilizar en el presente trabajo de investigación. Se optó por esta temática ya que cumple con los requisitos de extrapolación en el tratamiento de la información, todo esto complementado con bibliografía afín al tópico investigado.

Se optó por las tesis de ciencias forestales porque permite aplicar un criterio de descripción de la información más específico: la especie forestal con nombre científico y común, además de coordenadas geográficas (latitud y longitud). Este último criterio es posible de incorporar en otras colecciones del repositorio como *Geología*, *Geografía*, *Agronomía*, *Arquitectura*, entre otras.

Se realizó la recolección de la base de datos completa de registros del Repositorio Académico de la Universidad de Chile a través del sistema OAI-PMH, utilizando el software MarcEdit para MacOs. Este sistema permite la extracción de registros en formato Dublin Core XML.

Se realizó también un estudio acerca de las ontologías asociadas a sistemas bibliográficos, para lo cual se recogieron distintos modelos como: Europeana (**ver Anexo C**. *Modelo de Datos Europeana*) y British Library (**ver Anexo D**. *Modelo de Datos para libros de la British Library*), ambas han publicado sus modelamiento con las ontologías. Para saber de la selección de estas ontologías revise el punto **3.4** sobre **estudio de ontologías**.

Se realizó una investigación del estado del arte tanto bibliográfica, como la revisión de sitios extranjeros y nacionales, vinculados al área del uso de datos abiertos enlazados aplicado a sistemas bibliográficos, expuesto en el *capítulo 2* y en la bibliografía final de este estudio.

Para el desarrollo de la aplicación piloto se contactó a un experto en programación el cual desarrolló parte del código en lenguaje Php que permite realizar consultas al sistema.

3.1. Elección del área del conocimiento a trabajar

Para desarrollar el proyecto ha sido necesario definir una serie de características que debe reunir el set de datos a trabajar, entre los cuales se encuentran los siguientes:

- El set debe ser de acceso abierto.
- Debe cubrir una temática que permita extrapolar procedimientos en el tratamiento de información.
- Que posea el potencial de incorporar nuevos metadatos.
- Que la temática que sea de interés nacional.
- Que posea valor patrimonial para la universidad.
- Que sea factible de incorporar información georeferenciada.

A partir de estos criterios se optó por las tesis de Ingeniería Forestal como base, sin embargo por el alto volumen de trabajo que implicaba la limpieza y preparación de registros para el desarrollo del modelo se optó por trabajar con 25 registros, los cuales cuentan con las características base para representar este tipo de tesis. Esto debido a que al menos el 85 % de las tesis del área forestal cubren aspectos relacionados a coordenadas geográficas y un porcentaje por sobre el 95 % menciona especies forestales con su nombre científico. El set de datos elegido también es posible de extrapolar hacia temáticas como la Agronomía, donde las coordenadas geográficas vuelven a ser utilizadas así como el nombre científico también está presente en gran parte de esta colección dentro del repositorio; Arquitectura; así como Geografía, Geología o Minería.

3.2. Recolección de datos

Para la recolección de datos, se determinó trabajar con el Repositorio Académico de la Universidad de Chile. Además los datos recolectados cuentan con gran potencial de expansión para su descripción, el sistema que los almacena es Dspace, el que cuenta con un servicio de OAI a través del cual es posible extraer los metadatos estructurados como XML.

En cuanto a la recolección de registros se utilizó el programa computacional Marcedit, el cual permite realizar una recolección de una porción o el total de registros a través del sistema Open Archive Initiative, extrayendo los datos en formato XML tal como muestra la figura 3.1. En primera instancia se extrajo un total aproximado de 42.000 registros equivalente al total del Repositorio

Académico. Se optó por realizar una primera muestra de 252 registros del total de la colección de la Facultad de Ciencias Forestales, para finalmente trabajar una submuestra de 25 registros los cuales representan en un 90 % el tipo de información que cubre el área forestal, es decir, estudiar alguna especie forestal en algún punto del país. Utilizar este universo permite extrapolar la metodología al resto de la colección de forestal.

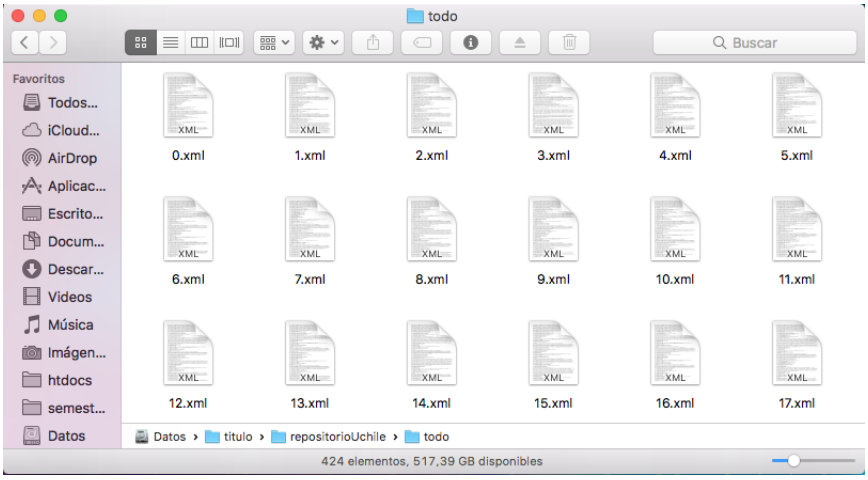


FIGURA 3.1. Archivos crudos, tal como los entrega Marcedit.

3.3. Descripción de los conjuntos de datos

El set de datos de 25 registros a los cuales se denominará como registro crudo, pues están tal como los entrega el sistema a través de la recolección. Se observó el potencial de esta información, la que viene estructurada en XML bajo la norma de Dublin Core cualificado o extendido tal como se aprecia en la figura 3.2. El potencial está en directa relación al tipo de contenido que se ha decidido trabajar, el cual al igual que otras colecciones dentro del repositorio Académico, el archivo en texto completo menciona el lugar geográfico donde se llevó a cabo el estudio descrito en la tesis, alojando este tipo de dato en el campo dc:subject. Pese a que escasamente registran esa información durante la descripción de la información.

FIGURA 3.2. Registro en XML, bajo norma Dublin Core que entrega MarcEdit.

3.4. Estudio de ontologías

Se realizó un estudio acerca de las ontologías a usar las cuales se detallan en el punto siguiente (**3.4.1**). Se tomó cómo modelo de datos a revisar la Biblioteca Británica el cual se aprecia en la figura D.1 y la Europeana, también posible de analizar a través de la figura C.1. Esto pues ambos sistemas han publicado sus estructuras de datos de manera abierta, lo cual permite su comprensión, facilitando además la reutilización de datos de que disponen y por otro lado la creación de una estructura de datos propia. Existen hasta ahora una serie de ontologías que permiten llegar a describir información bibliográfica de manera directa y en algunos casos simple, sin embargo con una sola ontología no se puede pretender realizar una descripción completa, según se explica en el punto **3.5**. Por otro lado existen usos que son importantes de rescatar para su reutilización. Es fundamental compartir no solo los datos, sino también las ontologías, para que se puedan reutilizar de manera más fácil. También permiten establecer vínculos entre distintas iniciativas como integrar datos de DBpedia en mi sistema. Es por esto que (Konstantinou, Spanos, et al., 2014) propone que: “la mera reutilización de los términos de ontología populares no conduce por sí sola a los verdaderos 5 estrellas de Linked Data (Berners-Lee, 2006). Para lograr esto último, las entidades y conceptos

referenciados en los metadatos del repositorio deben ser reconocidos y los identificadores adecuados para ellos deben ser encontrados entre los conjuntos de datos RDF ya publicados, para establecer vínculos entre ellos. Estas entidades incluyen principalmente autores, sujetos y palabras clave de los elementos almacenados en una biblioteca digital".

3.4.1. Entre las ontologías estudiadas se encuentran:

FOAF: es un proyecto dedicado a vincular personas e información a través de la Web. Según Segaran (2009) “FOAF se utiliza para representar información sobre personas, tales como sus nombres, cumpleaños, fotos, blogs y especialmente a las otras personas que conocen”. Por otra parte, FOAF describe el mundo usando ideas simples inspiradas en la Web. En las descripciones de FOAF, solo hay varios tipos de cosas y enlaces, que llamamos propiedades. Los tipos de las cosas de las que hablamos en FOAF se llaman clases. Por lo tanto, FOAF se define como un diccionario de términos, cada uno de los cuales es una clase o una propiedad. Otros proyectos junto a FOAF ofrecen otros conjuntos de clases y propiedades, muchas de las cuales están vinculadas con las definidas en FOAF. (Extraído desde el sitio `http://xmlns.com/foaf/spec/`)

bibo: la ontología bibliográfica bibo describe las cosas bibliográficas en la Web semántica en RDF. Esta ontología puede ser utilizada como una ontología de citas, como una ontología de clasificación de documentos, o simplemente como una forma de describir cualquier tipo de documento en RDF. Se ha inspirado en muchos formatos de metadatos de descripción de documentos existentes y puede utilizarse como base común para convertir otras fuentes de datos bibliográficos. La ontología bibliográfica es una aplicación del Marco de Descripción de Recursos (RDF) porque el área temática que estamos describiendo - citas y referencias bibliográficas - tiene tantos requerimientos competitivos que un formato independiente no los captura o llevaría a tratar de describir estos en un número de formatos incompatibles. Mediante el uso de RDF, la ontología bibliográfica gana un potente mecanismo de extensibilidad, que permite que las descripciones basadas en *bibliographic-ontology* se mezclen con las afirmaciones hechas en cualquier otro vocabulario RDF. (Extraído desde el sitio http://bibliontology.com/)

BIO: la ontología de BIO contiene términos útiles para descubrir más acerca de las personas y sus antecedentes se cruza con la información genealógica. El enfoque adoptado es describir la vida de una persona como una serie de eventos claves interconectados, alrededor de los cuales se puede tejer otra información. Es interesante de agregar pues nos permite definir por ejemplo relación

de parentesco o el puesto de trabajo o cargo público involucrado en un evento, en síntesis es un vocabulario para publicar información biográfica (Extraído desde el sitio http://vocab.org/bio/)

Schema: a principios de junio de 2011, los tres grandes motores de búsqueda Bing, Google y Yahoo! introdujeron Schema.org, una colección de términos que los *webmasters* pueden usar para marcar sus páginas para mejorar la visualización de los resultados de búsqueda. Este sitio es un esfuerzo complementario de personas de la comunidad de datos enlazados para apoyar el despliegue y el uso de Schema.org con un enfoque especial en Linked Data. (Extraido desde el sitio http://schema.rdfs.org/)

MADS/RDF: es un sistema de organización del conocimiento (KOS) diseñado para su uso con valores controlados para nombres (personales, corporativos, geográficos, etc.), tesauros, taxonomías, sistemas de encabezamiento de temas y otras listas de valores controlados. Está estrechamente relacionado con SKOS, el Sistema de Organización Simple del Conocimiento y un vocabulario ampliamente respaldado y adoptado por RDF. A diferencia de SKOS, MADS/RDF está diseñado específicamente para soportar datos de autoridad utilizados y necesarios en la comunidad LIS y sus sistemas tecnológicos. Esta comunidad se refiere a bibliotecas, archivos, museos u otro tipo de instituciones del ámbito cultural. Por ejemplo, MADS/RDF proporciona un medio para registrar datos del formato de Autoridades de Catalogación de Lectura Automática (MARC) en RDF para su uso en aplicaciones semánticas y proyectos de Datos Vinculados.(Extraído desde el sitio http://www.loc.gov/standards/mads/rdf/)

Dublin Core: es un conjunto de metadatos moderadamente pequeño el cual se dividide en dos vocabularios: *dc elements* cuenta con un espacio de nombres en http://purl.org/dc/elements/1.1/ y *dc terms* cuenta con su espacio de nombres en http://purl.org/dc/terms/. Los elementos dc contienen 15 propiedades expuestas en la tabla 1.1. Los términos de *dc terms* contienen 22 clases y 55 propiedades explicadas en la tabla B.1. No es una ontología OWL 2 DL porque no se basa en absoluto en las construcciones OWL. (Extraído desde el sitio https://www.w3.org/wiki/Good_Ontologies) Dublin Core es un modelo de metadatos elaborado y auspiciado por la Dublin Core Metadata Initiative, una organización dedicada a fomentar la adopción extensa de los estándares interoperables de los metadatos y a promover el desarrollo de los vocabularios especializados de metadatos para describir recursos que permitan sistemas más inteligentes en el descubrimiento del recurso. (Extraído desde el sitio https://es.wikipedia.org/wiki/Dublin_Core).

GeoNames: la Ontología de GeoNames hace posible agregar información semántica geoespacial a Word Wide Web. Más de 11 millones de topónimos GeoNames ahora tienen una URL única con un servicio Web RDF correspondiente. Otros servicios describen la relación entre los topónimos. (Extraído desde el sitio http://www.geonames.org/)

SKOS: es un modelo común de datos para compartir y vincular sistemas de organización del conocimiento a través de la Web Semántica. (Extraído desde el sitio https://www.w3.org/2009/08/SKOS-reference/SKOS.html). SKOS (siglas de Simple *Knowledge Organization System*) es una iniciativa del W3C[1] en forma de aplicación de RDF que proporciona un modelo para representar la estructura básica y el contenido de esquemas conceptuales como listas de encabezamientos de materia [2], taxonomías[3], esquemas de clasificación[4], tesauros [5] y cualquier tipo de vocabulario controlado. (Extraído desde el sitio https://es.wikipedia.org/wiki/Simple_Knowledge_Organization_System)

WSG84: un vocabulario para representar información de latitud, longitud y altitud en el dato de referencia geodésico. (Extraído desde el sitio http://lov.okfn.org/dataset/lov/vocabs/geo)

3.5. Elección de la ontología a utilizar

Las ontologías están descritas brevemente en el punto anterior. ¿Por qué se debe utilizar más de una ontología? Porque una sola ontología no puede describir el universo de información que se posee. Todas ellas reunidas permiten crear una estructura más compleja, individuos que se componen por: *object properties* que me permiten vincular los datos que poseen con otras fuentes de información estructuradas semánticamente y *data properties*, las cuales contienen un valor textual o numérico dependiendo de la propiedad. Por ejemplo, al describir información georeferenciada, que describe un punto específico en un plano definido por latitud y longitud, la ontología bibo no cuenta con una propiedad que pueda expresar o representar este tipo de valor, por lo tanto se utilizan otras ontologías como GeoNames y WGS84. Ambas se complementan para describir un punto específico del plano, utilizando wgs84:lat para latitud y wgs84:long para longitud, junto con geonames:name para cubrir todo el conjunto de propiedades específicas para este área. Tal y como plantea (Konstantinou, Spanos, et al., 2014): “además, un conjunto de datos RDF que utiliza términos de

[1] https://es.wikipedia.org/wiki/W3C
[2] https://es.wikipedia.org/wiki/Encabezamientos_de_materia
[3] https://es.wikipedia.org/wiki/Taxonom%C3%ADa
[4] https://es.wikipedia.org/wiki/Clasificaci%C3%B3n
[5] https://es.wikipedia.org/wiki/Tesauro

vocabularios ampliamente difundidos tiene una probabilidad mayor de atraer referencias de terceros, en comparación con un conjunto de datos basado en ontologías hechas a medida cuyo propósito exacto puede no estar claro".

3.6. Vocabulario controlado en datos abiertos enlazados

De acuerdo a la norma (NISO, 2010) el vocabulario controlado se define como: “el que se utiliza para mejorar la eficacia de almacenamiento y recuperación en sistemas de información, sistemas de navegación Web, y otros entornos que tratan de identificar y localizar el contenido deseado a través de algún tipo de descripción usando el lenguaje". De acuerdo a (Castillo Guerrero, 2008) “el objetivo principal de control de vocabulario es lograr la coherencia en la descripción de objetos y el contenido para facilitar la recuperación".

El uso de vocabulario controlado según (Radio y Hanrath, 2016) radica en que: “un vocabulario controlado o un archivo de autoridad de nombre, mejoraría la consistencia de los datos en Universidad de Kansas". El mismo autor expresa: “la introducción de un vocabulario controlado podría ayudar a asegurar que este tipo de trabajo de corrección no tenga que ser repetido". Un objetivo secundario es la posibilidad de un descubrimiento mejorado a través de la exposición de los registros como Linked Data. Pues es muy usual encontrar inconsistencia de materias o autores en sistemas de repositorios además de las sabidas falta de ingreso de información.

Para el presente proyecto se desarrollaron dos instancias para la creación de vocabulario controlado:

- Creación del RDF SKOS-subject. (ver tabla 4.3)
- Creación del RDF Autor-Authority (ver tabla 4.6)

Cada uno de estos grafos permiten la estructuración tanto de descriptores bajo el concepto de autoridad como a su vez las personas involucradas en una tesis.

4. APLICACIONES A UN CASO DE ESTUDIO

La idea del proyecto es definir un procedimiento que pueda entregar una visión general del camino a seguir para la obtención de los archivos RDF. Esto resulta útil a la hora de planificar un procedimiento de datos abiertos enlazados, pues como se muestra en la 4.1 es posible comprender que ciertas etapas como la limpieza de datos requiere de un mayor tiempo que otros hitos. Por otro lado, es fundamental comprender que la generación de archivos RDF en el punto 6 será junto a la reconciliación en el punto 7 y que ambos procesos serán iterativos. Esto ya que mientras los datos dispuestos en el punto 6 comiencen a encontrar respuesta en fuentes externas como internas, se podrán ir enriqueciendo los registros bibliográficos de la información con definiciones o imágenes, elementos que de partida no se poseen. Lo interesante de un proceso de datos vinculados en la Web es que es definitivamente dinámico, por lo tanto, es posible iniciar un proceso de aprendizaje junto a la información que se va relacionando, descubriendo vínculos que no son necesariamente obvios.

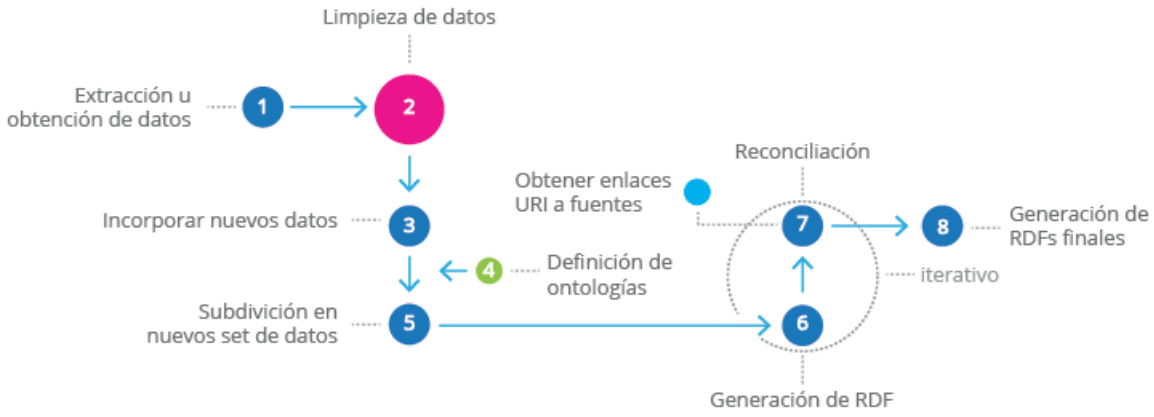


FIGURA 4.1. Procedimiento para generar RDF.

4.1. ETL, uso de OpenRefine

Para desarrollar el presente proyecto se utilizó OpenRefine, herramienta *open source*, que permite realizar una gran cantidad de tareas como: limpieza, transformación, ampliación de datos y una serie de acciones que le dan a este programa un funcionamiento para trabajar con datos y modelar un archivo de salida (como RDF a través del complemento RDF extension 0.8).

Para utilizar esta solución es necesario tener instalado un entorno de trabajo con Java SE Development Kit, última versión. Posteriormente, es necesario descargar la versión OpenRefine 2.5, en este caso se utilizó la versión para Mac OSX[1].

4.2. Carga de datos

OpenRefine puede cargar distintos tipos de archivo entre ellos: TSV, CSV, +SV, Excel (xls y xlsx), JSON, XML, RDF como XML e incluso conectarse a hojas de cálculo de Google tal y como lo muestra la figura 4.2.

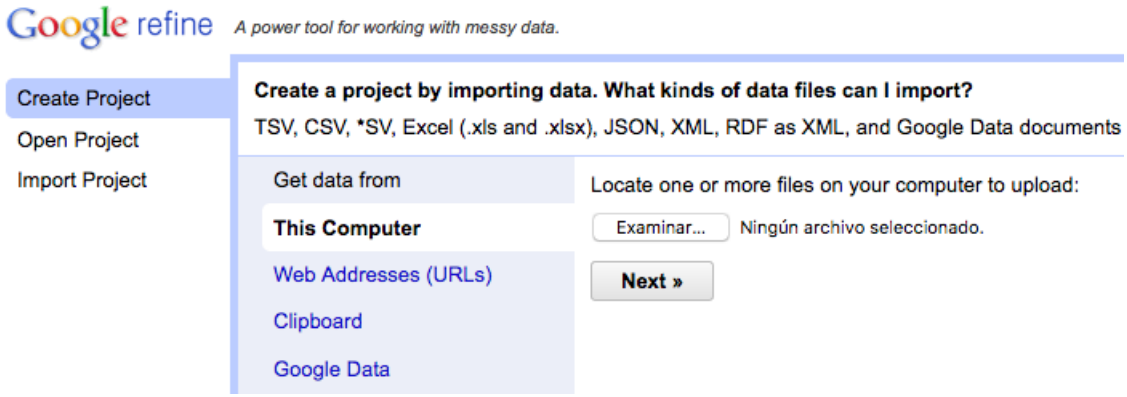


FIGURA 4.2. Interfaz de OpenRefine para cargar archivos.

Al seleccionar el archivo a trabajar, el sistema solicitará definir como estructurará la información para ser desplegada en el ambiente de OpenRefine (parseo).

4.3. Limpieza de datos

No siempre los datos vienen listos para utilizar, es posible que falte información o no esté correctamente ingresada. En el presente proyecto se procedió en primera instancia a cargar tres archivos xml con un total de 252 registros correspondientes a la colección de la Facultad de Ciencias Forestales, no obstante se analizaron 25 registros correspondiente a las primeras tesis obtenidas por la exportación, no se trabajaron más que la cantidad mencionada por el tiempo que involucra la adecuación de metadatos, además que los registros en general utilizan los mismos datos base (autor, título y materia), muestra suficiente para desarrollar este proyecto.

[1] `http://openrefine.org/download.html`

Se eliminan las columnas que no aportan información relevante como: *datestamp* dato que corresponde a la fecha de importación, *dc:source*, *header - status*, *resumptionToken - cursor*, *resumptionToken-completeListSize*, *ListRecords-resumptionToken* columnas que se encuentran sin información.

Para este proyecto se procedió a renombrar las columnas que contienen los datos para facilitar su lectura y la posterior manipulación como se muestra en la tabla 4.1.

Nombre original entregado por la fusión	Nombre definitivo
metadata - oai_dc:dc - dc:subject	dc:subject
metadata - oai_dc:dc - dc:date	dc:date
metadata - oai_dc:dc - dc:creator	dc:creator
metadata - oai_dc:dc - dc:contributor	dc:contributor
metadata - oai_dc:dc - dc:identifier	dc:identifier
metadata - oai_dc:dc - dc:description	dc:description
metadata - oai_dc:dc - dc:title	dc:title
metadata - oai_dc:dc - dc:rights	dc:rights
metadata - oai_dc:dc - dc:type	dc:type
metadata - oai_dc:dc - dc:language	dc:language
metadata - oai_dc:dc - dc:publisher	dc:publisher

TABLA 4.1. Renombre de columnas equivalente a los campos obtenidos

Se utilizó la opción de crear facetas en OpenRefine por las distintas columnas cargadas para leer y detectar inconsistencias como las expuestas en la figura 4.3 y 4.4.

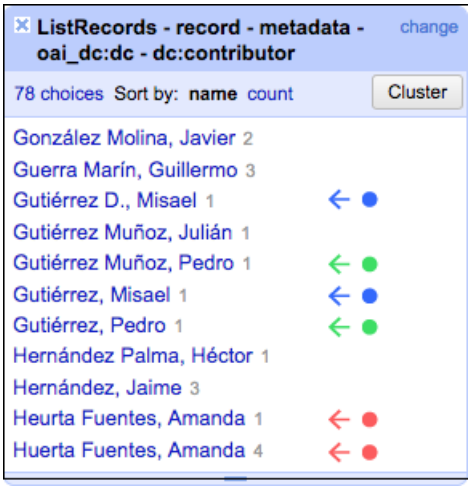


FIGURA 4.3. Errores comunes detectados con facetas.

FIGURA 4.4. faceta de materias, antes de limpiar.

Para **dc:date** que contenía la fecha de ingresó del registro al sistema como un campo repetible, el cual contiene la fecha de publicación teniendo estos significados muy distintos, ante lo cual se dejó sólo esta última, tal y como se aprecia en la figura 4.5.

El Campo **dc:language**, se decidió utilizar una forma más comprensible para describir el idioma, por lo que se cambió la abreviatura es por español, tal como muestra la figura 4.6.

dc:date
2012-09-12T18:22:49Z
2012-09-12T18:22:49Z
2008

FIGURA 4.5. Limpieza de información.

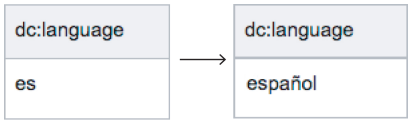


FIGURA 4.6. Definición de políticas.

Cualquier campo que tuviese datos repetibles se optó por pasarlos desde una columna a una sola fila. Esto facilitaría más adelante la estructura del archivo RDF tal como se muestran las figuras 4.7 y 4.8.

dc:subject
Ingeniería Forestal
Peumo
Quillay
Riego
Sequías
Cryptocarya alba
Quillaja saponaria

FIGURA 4.7. Estado original del campo subject.

dc:subject	dc:subject2	dc:subject3	dc:subject4	dc:subject5	dc:subject6	dc:subject7
Ingeniería Forestal	Peumo	Quillay	Riego	Sequías	Cryptocarya alba	Quillaja saponaria

FIGURA 4.8. Resultado de trasposición.

La limpieza de datos entregó como resultado final, un primer conjunto de datos, al cual se le denominó “recursos”. Este conjunto incluía los siguientes metadatos:

- dc:subject
- dc:date
- dc:creator
- dc:contributor
- dc:identifier
- dc:description
- dc:title
- dc:rights
- dc:type
- dc:language
- dc:publisher
- dc:abstract

4.4. Incorporación de metadatos

Para el desarrollo de este ejercicio se optó por incorporar una serie de campos adicionales que no existían o no se habían utilizado en el repositorio. Esto permitirá potenciar aún más la experiencia de búsqueda y obtención de información por parte del usuario final. Entre los campos agregados se encuentran los siguientes:

Resumen: este campo a pesar de estar presente en el Repositorio Académico es opcional, en las tesis de pregrado de forestal se ha llenado en 14 de 162 registros. Es necesario hacer hincapié que todas las tesis deben presentar resumen como parte de su formato obligatorio por lo que se sabe que ésta información está necesariamente presente en los archivos que acompañan al registro, pues forma parte de la tesis. En el caso de este estudio se procedió a revisar todos los archivos PDF de las tesis con el fin de extraer esta información. Se incorpora al campo dc:abstract pues acá se puede encontrar información como la ubicación geográfica además del nombre de la especie forestal en estudio, entre otra información altamente relevante.

Coordenadas geográficas: las coordenadas geográficas no son parte actualmente del sistema bibliográfico para descripción de información, sin embargo al menos un 85 % de las tesis de pregrado en el área forestal mencionan algún lugar donde se realizó el estudio. Aproximadamente sólo 15 de 162 tesis descritas en el Repositorio Académico, mencionan el lugar geográfico y en términos muy genéricos, utilizando conceptos como Chile, Chile Central, Cordillera de la Costa, etc., en circunstancias que dichas tesis sí nombran lugares específicos como Lolol, Fundo Bagual, etc. Para suplir esta falta de información se decidió buscar y agregar el lugar mencionado en la tesis, junto con sus coordenadas geográficas de latitud y longitud al nuevo set de datos. Para incorporar las coordenadas se utilizó OpenRefine, incluyendo la API de Google Maps para saber exactamente la posición geográfica del estudio a describir esto se muestra en la figura 4.9. Luego se aprecia el resultado del requerimiento con una notación JSON en la figura 4.10. Finalmente se observa el código necesario para dejar sólo la información que se utilizará, en ese caso la longitud figura 4.11. Como suele suceder en un sistema que utiliza coordenadas geográficas, éstas se suelen repetir, por lo tanto se optó por crear dos conjuntos de datos para lugares: primero se creó uno que fuese capaz de cubrir todas las comunas o localidades mencionadas en las tesis (ver figura 4.4), después se creó otro subconjunto de datos que refleje la subdivisión administrativa mayor (ver figura 4.5), por ejemplo: Lolol pertenece a la Sexta Región del General Bernardo O'Higgins. Esto permitiría establecer relaciones por regiones. Esta información se incorpora con el nombre *place*, y puede ser repetible.

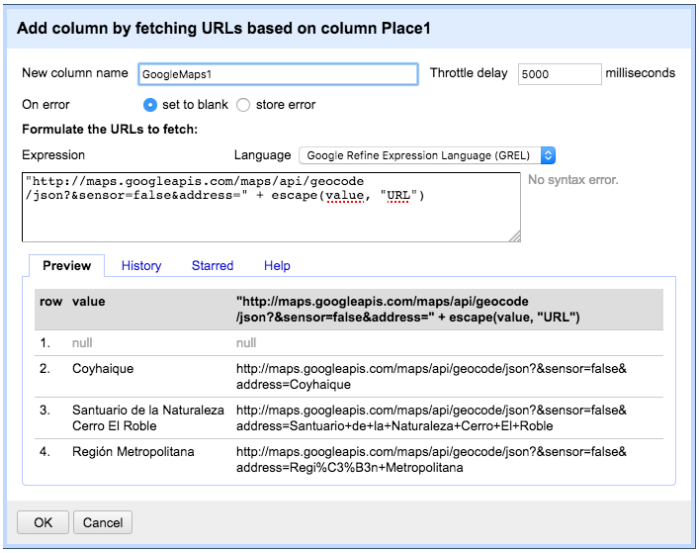


FIGURA 4.9. Utilización de Google Maps para obtener coordenadas geográficas.

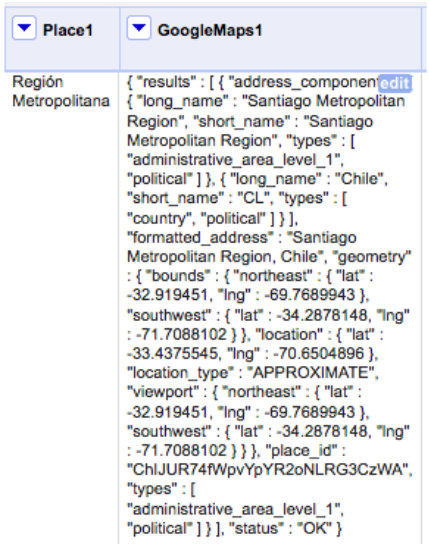


FIGURA 4.10. Resultado como lo entrega Google Maps en formato JSON.

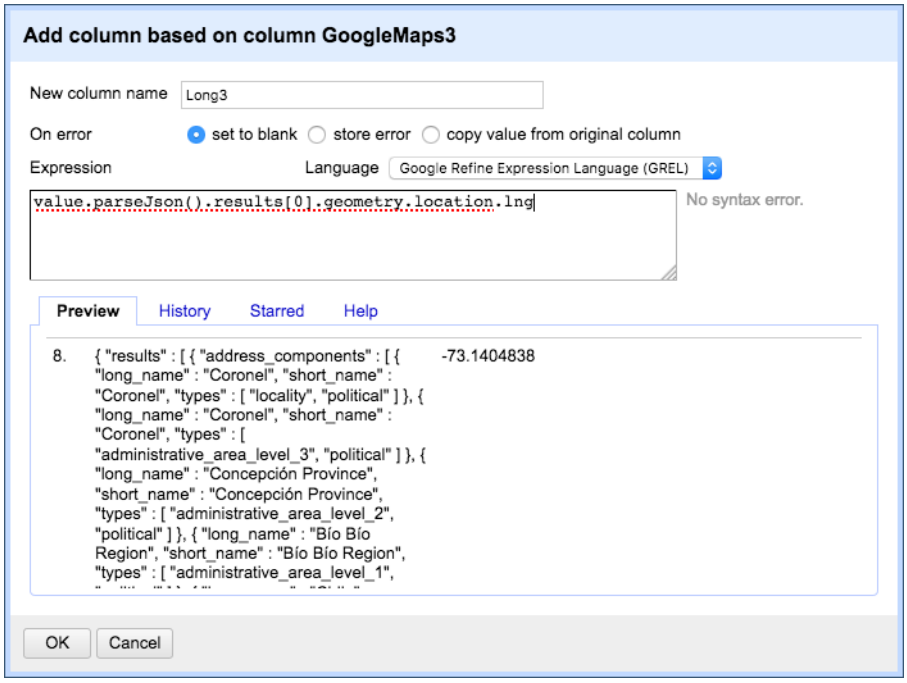


FIGURA 4.11. Transformación del JSON de Google Maps a solo Longitud.

Nombres científicos: si bien el nombre científico se incorpora a través de las materias en un registro, esto no siempre sucede, por lo que ha sido necesario incorporar todos los nombres faltantes para el conjunto de datos. Esta incorporación permitirá realizar un proceso denominado reconciliación con alguna fuente de datos como DBpedia en español. Se considera esta fuente pues como expresa (Auer, Lehmann, y Hellmann, 2009): "un gran conjunto de datos de referencia que proporciona conocimientos enciclopédicos sobre una multitud de dominios diferentes ya está disponible". Se incorpora en el campo nombre-científico y puede ser repetible.

Reasignación de materias: se decidió utilizar un sistema basado en AGROVOC[2] para la asignación de materias. Esto aportaría consistencia a la información, pues se detectó la utilización de conceptos genéricos o insuficientes para describir el contenido. También fue posible observar inconsistencia ya que se utilizan encabezamientos de materia junto con palabras claves, lo que impide relacionar registros similares. Fueron reasignados solo descriptores para toda la muestra. Para garantizar que el sistema funcione se optó por crear otro set de datos complementario, para la creación de un nuevo grafo RDF, el cual se relacionaría con el set de datos principal. Esto para que cumpla con el modelo de autoridad de materia y evitar la errónea utilización de conceptos. También puede ser repetible y se definió como política utilizar como máximo cinco descriptores.

Creator (Autor): se procedió a crear un sistema propio de autoridades de autor (personas) que entregase coherencia a la utilización de nombres de autores, ya que fue posible detectar autores iguales, escritos de distinta forma o con errores de escritura. Para evitar este tipo de inconsistencia se creó un nuevo set de datos. Es necesario mencionar que este subset de datos sirve para validar a alumnos como académicos, por lo tanto se relaciona con el campo dc:creator y dc:contributor el cual puede ser repetible.

URI: se decidió agregar este campo para que facilitara la identificación de individuos y sus propiedades, esta cadena de caracteres es única para el recurso. Esto se explica en el punto 5.3.

4.5. Creación de nuevos conjuntos de datos

Después de limpiar y agregar nuevos metadatos, se procedió a dividir el set de datos principal denominado recursos en los siguientes conjuntos de datos:

[2] `http://aims.fao.org/es/agrovoc`

A **Recursos** se le agregó nuevos campos marcados en color, además quedó compuesto por lo expuesto en la siguiente tabla:

Campo	Definición
URI	Identifica las obras de la red de forma unívoca
nombre-cientifico1	Nombre científico de la especie forestal
nombre-cientifico2	Nombre científico de la especie forestal
nombre-cientifico3	Nombre científico de la especie forestal
nombre-comun1	Nombre común conocido de la especie forestal
nombre-comun2	Nombre común conocido de la especie forestal
nombre-comun3	Nombre común conocido de la especie forestal
nombre-no-autorizado	Nombre común no autorizado
place1	Lugar del estudio
place2	Lugar del estudio
place3	Lugar del estudio
place4	Lugar del estudio
place5	Lugar del estudio
place6	Lugar del estudio
place7	Lugar del estudio
place8	Lugar del estudio
place9	Lugar del estudio
place10	Lugar del estudio
dc:subject1	Descriptor que expresa de qué trata la obra
dc:subject2	Descriptor que expresa de qué trata la obra
dc:subject3	Descriptor que expresa de qué trata la obra
dc:subject4	Descriptor que expresa de qué trata la obra
dc:subject5	Descriptor que expresa de qué trata la obra
dc:date	Fecha de publicación de la obra
SKOS:notation	Será el ID para la fuente externa DPbedia
dc:creator	Autor de la obra
dc:contributor1	Contribuidor de la obra
dc:contributor2	Contribuidor de la obra
dc:contributor3	Contribuidor de la obra
dc:contributor4	Contribuidor de la obra
dc:contrubutor5	Contribuidor de la obra
dc:identifier	Identificador de la obra heredado de la exportación
dc:description	Descripción de la obra
dc:abstract	Resumen de la obra
dc:title	Título de la obra
dc:rights	Derechos de la obra, heredado de la exportación
dc:type	Tipo de recurso
dc:language	Lenguaje de la obra
dc:publisher	Publicador de la obra

TABLA 4.2. Campos agregados después de la limpieza de datos

SKOS-subject: en este caso se procedió a crear un nuevo grafo para el control de descriptores, el cual pudiese primero distinguir un término de otro a través de la URI. Esto nos permite contar con un concepto autorizado para cada registro a través del SKOS:prefLabel sino que además este grafo debía contener una estructura básica para cada descriptor, incluyendo la fuente de donde se extrae la información, así por lo tanto este set queda compuesto por las siguientes propiedades:

Campo	Definición
URI	Identifica los descriptores de la red de forma unívoca
SKOS:prefLabel	Descriptor preferido o autorizado
enlaceAgrovoc	Enlace a definición del descriptor en AGROVOC
SKOS:broader	Descriptor más amplio utilizado
SKOS:altLabel	Término no autorizado
SKOS:narrow	Descriptor específico
source	Fuente de información de donde se obtiene la información del descriptor

TABLA 4.3. SKOS-subject - Para descriptores

Places: como grafo debe incluir una propiedad que permita identificar el lugar de manera unívoca por lo cual incluye también un URI, una propiedad como toponymName el cual represente la forma autorizada para el lugar que se hace referencia. Se incluye la latitud y longitud como parte de las coordenadas geográficas, además del nombre oficial de la región por lo tanto el grafo está compuesto por las siguientes propiedades:

Campo	Definición
URI	Identifica los lugares de la red de forma unívoca
toponymName	Nombre de lugar autorizado
lat	Latitud
long	Longitud
region	Nombre oficial de la región

TABLA 4.4. Places - Para lugares geográficos

PlacesRegiones: se incorpora una propiedad para URI así identificar las regiones de manera unívoca en el grafo, por otro lado, es posible encontrar la propiedad nomRegion para nombres oficiales de esta subdivisión política, se incluye el nombre de la región incluyendo el número en romano y finalmente el nombre coloquial para esta subdivisión. Este grafo se vincula con el grafo *place* e incluye los siguientes campos:

Campo	Definición
URI	Identifica las regiones de la red de forma unívoca
nomRegion	Nombre oficial de la región
Region	Nombre de región con número en romano
regionNombreColoquial	Nombre coloquial de la región, término no autorizado

TABLA 4.5. Identificación de las regiones de Chile

Author-Authority: este grafo permite mantener un control sobre los nombres de las personas que trabajan en relación a una tesis, ya sea como alumno tesista, o como profesor guía u otro cargo, para lo cual cuenta con URI para identificar unívocamente a la persona en el sistema. Luego es posible contar con una propiedad que indica la fuente de donde se obtuvo la información. También es parte de este grafo *foaf:name* para registrar el nombre de la persona. Además se incluyen formas alternativas para el nombre así como la profesión si es que está disponible la información. Las propiedades que componen este grafo son:

Campo	Definición
URI	Identifica los las regiones de la red de forma unívoca
source	Fuente de donde se obtiene la información
rdf:type	Define el tipo de recurso
uchile:relacionUchile	Si existe vínculo con la Universidad de Chile
foaf:name	Nombre de la persona
rdfs:label	Nombre de la persona, propiedad alternativa
schema:birthDate	Fecha de nacimiento
schema:deathDate	Fecha de muerte
foaf:member	Si pertenece a alguna unidad en la universidad
nombre alternativo	Si posee nombre alternativo por el cual sea conocido
nombre alternativo 2	Si posee nombre alternativo por el cual sea conocido
profesion	Profesión del titular del registro
madsrdf:hasExactExternalAuthority	Se obtiene una URI para acceder a autoridad de LOC[3]

TABLA 4.6. Definición de Personas

Reconciliación: esta acción permite buscar y relacionar términos del set de datos con fuentes externas, siempre y cuando la fuente cuente con un servicio de Sparql endpoint como: DBpedia, Biblioteca del Congreso Nacional en Chile, AGROVOC, entre otras. También es necesario reconciliar entre sí los distintos set de datos que se crearon para este proyecto los cuales son posibles de gestionar como archivos RDF. Este paso es fundamental para enriquecer y agregar valor a los datos que se poseen. Como a su vez, relacionar los grafos que se obtuvieron a partir del set de datos. Este proceso puede ser iterativo, con el fin de detectar inconsistencias o falta de relaciones claves. En las siguientes figuras se muestra el procedimiento que ofrece OpenRefine.

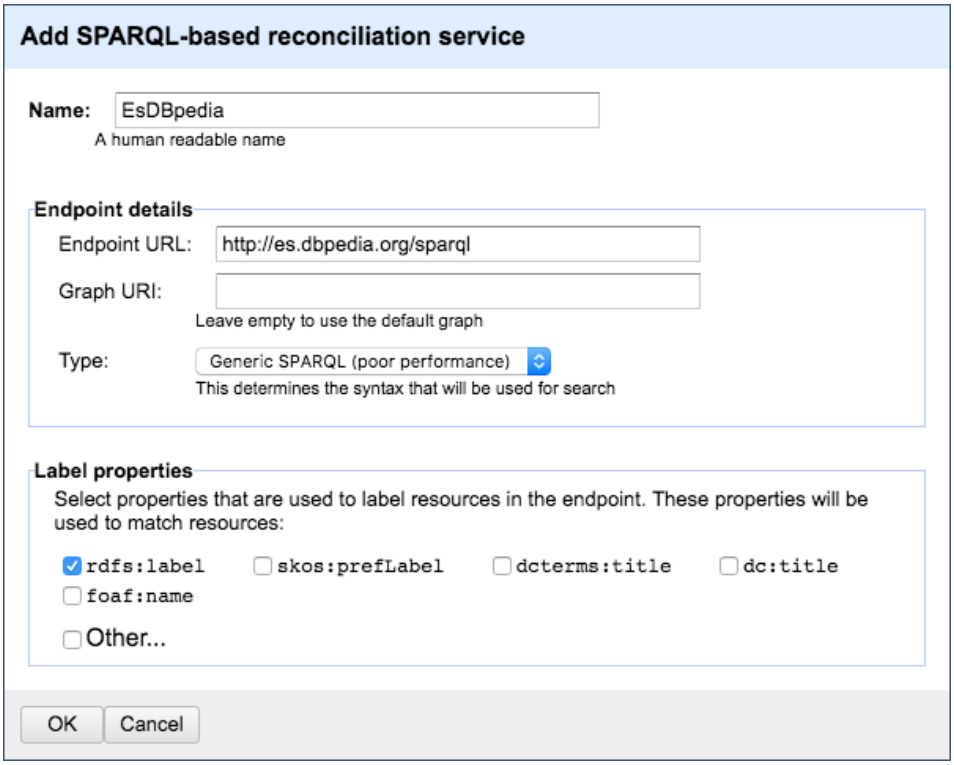


FIGURA 4.12. Se registra la dirección del endpoint para consultar.

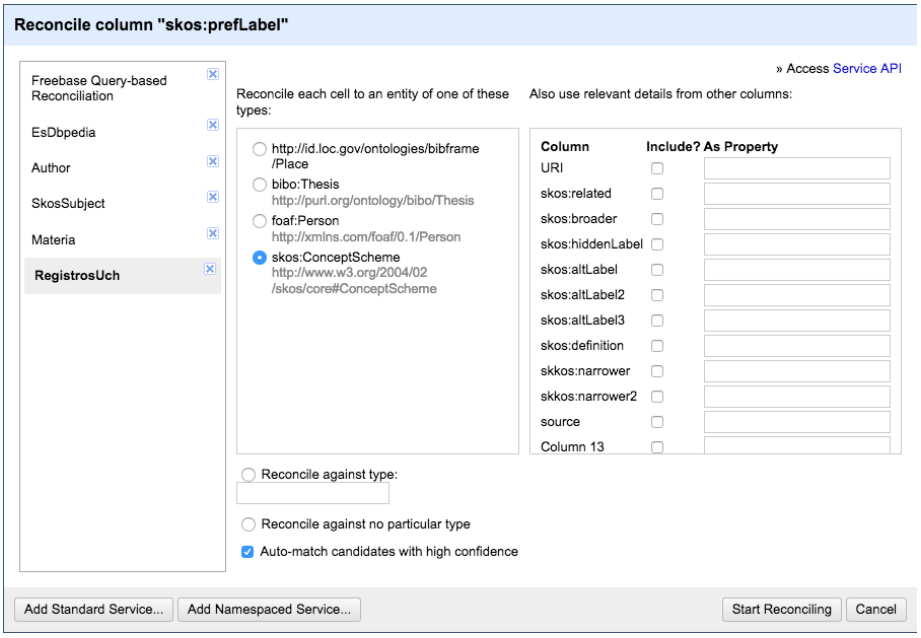


FIGURA 4.13. Proceso de reconciliación de OpenRefine.

4.5.1. Obteniendo la URI de fuente externa

Procedimiento de OpenRefine para extracción de URI externa.

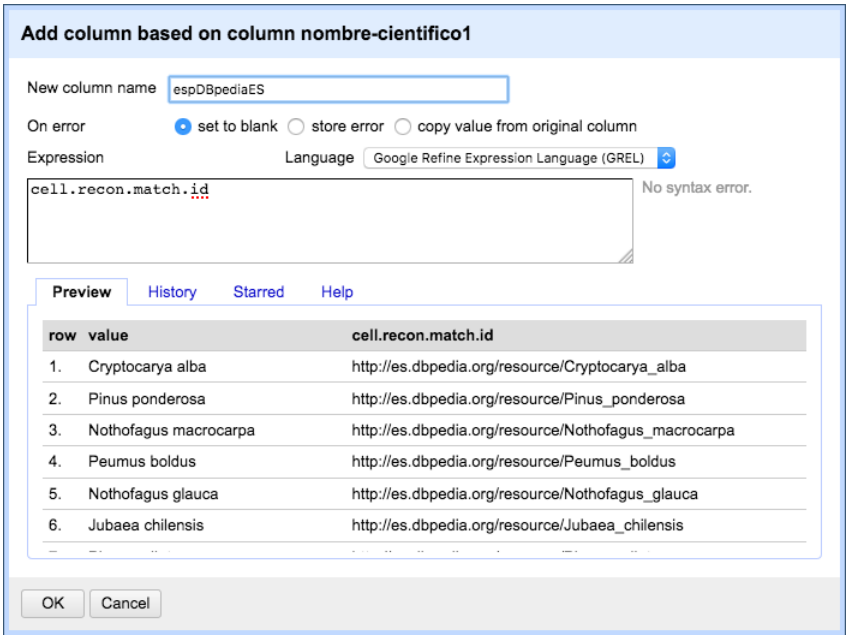


FIGURA 4.14. Extracción del URI que accede a fuente externa.

4.6. Generación de archivos RDF

OpenRefine permite a través de un complemento denominado RDF extension en su versión 0.8, estructurar el contenido que se encuentra en el modelo de datos de la figura 4.15. A continuación se mostrará paso a paso como se van tratando los datos para la generación de los archivos RDF, figura 4.16. Luego se muestra como agregar nuevos vocabularios para describir información en la figura 4.17.

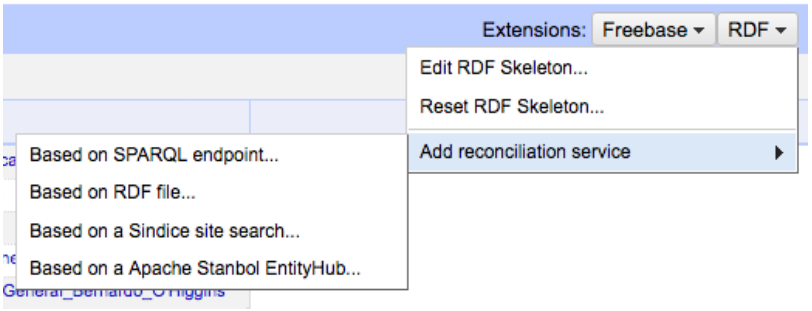


FIGURA 4.15. A través de Edit RDF Skeleton se genera RDF.

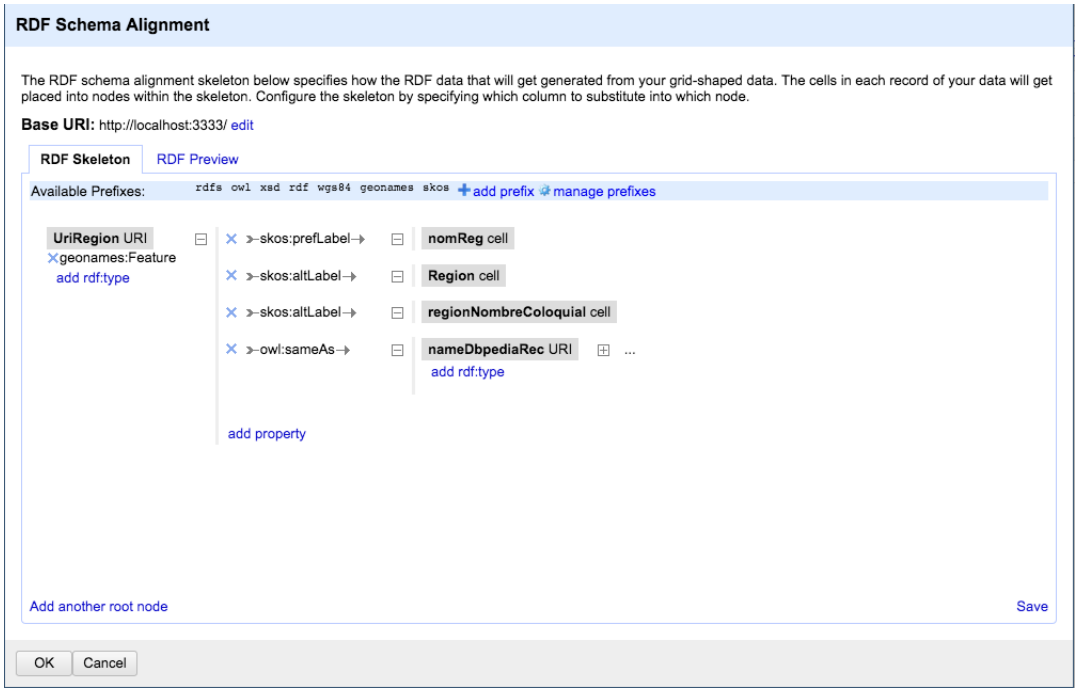


FIGURA 4.16. Interface de RDF extension.

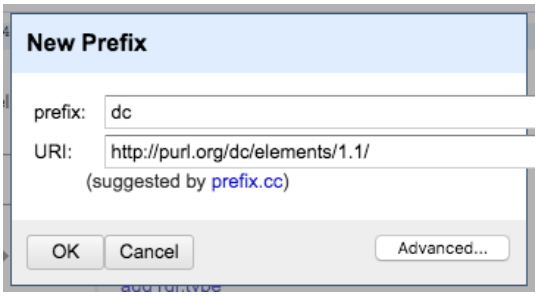


FIGURA 4.17. Agregando nuevos prefijos, equivalentes a nuevas ontologías.

OpenRefine permite exportar la información en distintos formatos, entre ellos RDF como RDF/XML o como RDF del tipo turtle.

5. RESULTADOS Y ANÁLISIS

5.1. Modelo de datos del proyecto

Al procesar los cinco conjuntos de datos es posible obtener un archivo RDF por cada uno, lo cual implica poseer cinco grafos que se interconectarán una vez dispuestos en OpenLink Virtuoso tal y como se muestra en la figura 5.1.

Para facilitar la comprensión de este punto se ha desarrollado un modelo de datos del sistema, donde es posible reconocer con cada una de las propiedades de los distintos grafos.

Es importante agregar en esta etapa todas las ontologías que utilizará el sistema, tal como: dc, SKOS, bibo, dcterms, Schema, FOAF, GeoNames, WGS84, BIO, se incorporan a OpenRefine.

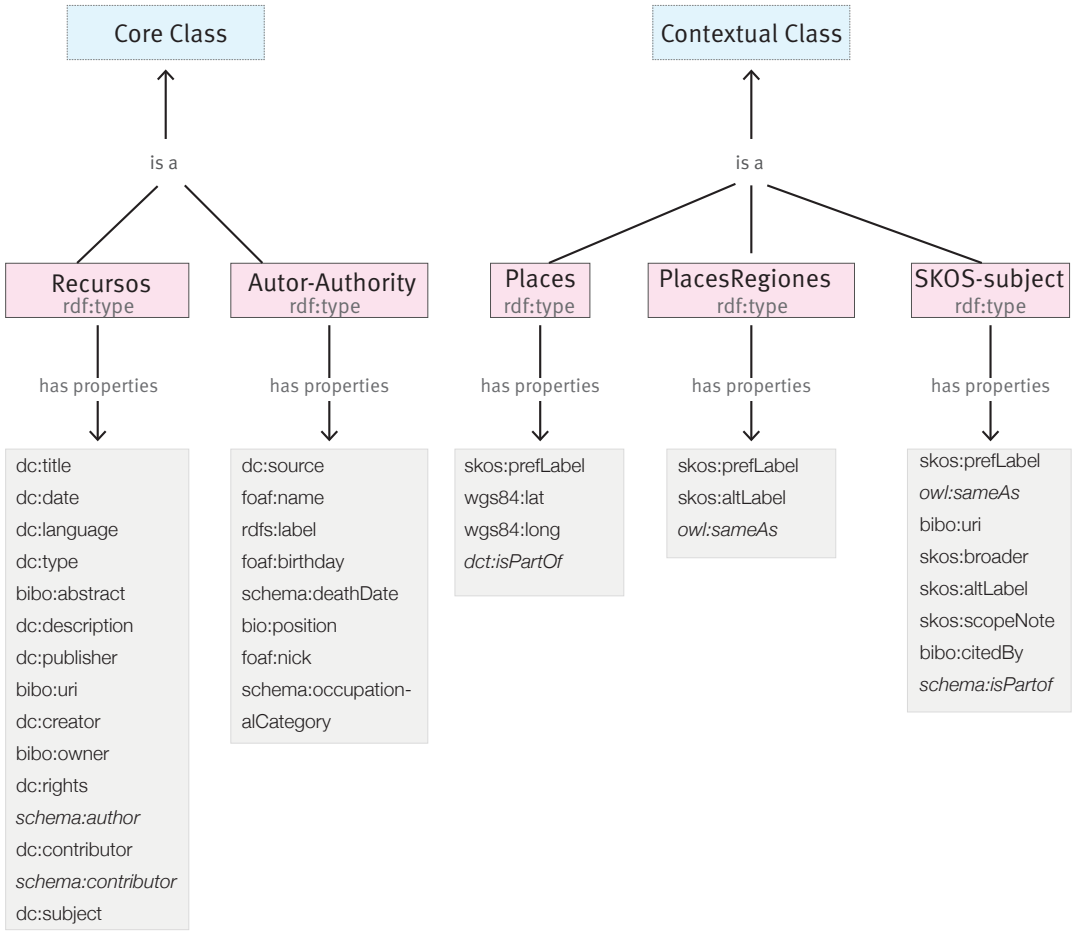


FIGURA 5.1. Modelo de datos propuesto por este proyecto.

5.2. Mapeo de metadatos a propiedades LOD

Una vez desarrollado el modelo de datos para el proyecto es fundamental realizar un mapeo entre los metadatos y las ontologías seleccionadas así será posible tener claridad de cómo serán estructurados los grafos RDF para cada conjunto de datos.

5.2.1. Mapeo de recursos

Nombre del metadato	Predicado	Tipo de objeto
nombre-cientifico	dc:subject	URI
IdDbpedia	SKOS:notation	Text
nombre-comun	dc:subject	URI
nombre-no-autorizado	dc:subject	URI
Place	geoname:name	Text
PlaceRec	dc:coverage	URI
dc:subjectRec	dc:subject	URI
dc:date	dc:date	Text
dc:creator	dc:creator	Text
dc:creator	bibo:owner	Text
EnlaceAutoridad	schema:author	URI
dc:contributor	dc:contributor	Text
DbpediaContr	schema:contributor	URI
dc:identifier	bibo:uri	URI
dc:description	dc:description	Text
dc:abstract	bibo:abstract	Text
dc:title	dc:title	Text
dc:rights	dc:rights	Text
dc:type	dc:type	Text
dc:language	dc:language	Text
dc:publisher	dc:publisher	Text

TABLA 5.1. Mapeo de Metadatos de recursos

5.2.2. Mapeo de metadatos SKOS-subject

Nombre del metadato	Predicado	Tipo de objeto
SKOS:prefLabel	SKOS:prefLabel	Text
uriAgrovocRec	owl:sameAs	URI
Column 13	bibo:uri	URI
SKOS:related	SKOS:related	Text
SKOS:altLabel	SKOS:altLabel	Text
SKOS:definition	SKOS:scopeNote	Text
source	bibo:citedBy	URI

TABLA 5.2. Mapeo de materias

5.2.3. Mapeo de metadatos Places

Nombre del metadato	Predicado	Tipo de objeto
toponymName	SKOS:PrefLabel	Text
dbpediaRec	owl:sameAs	URI
BnCRec	owl:sameAs	URI
lat	wgs84:lat	Text
lng	wgs84:long	Text
UriRegion	dct:isPartOf	URI
Region2Rec	dct:isPartOf	URI

TABLA 5.3. Mapeo de metadatos de lugares (Places)

5.2.4. Mapeo de datos PlacesRegiones

Nombre del metadato	Predicado	Tipo de objeto
nomReg	SKOS:PrefLabel	Text
Region	SKOS:altLabel	Text
regionNombreColoquial	SKOS:altLabel	Text
nameDbpediaRec	owl:sameAs	URI

TABLA 5.4. Mapeo de metadatos de regiones (PlacesRegiones)

5.2.5. Mapeo de metadatos de Author-Authority

Nombre del metadato	Predicado	Tipo de objeto
source	dc:source	URI
rdf:type	rdf:type	URI
uchile:relacionUChile	bio:position	Text
foaf:name	foaf:name	Text
rdfs:label	rdfs:label	Text
schema:birthDate	foaf:birthdate	Text
schema:deathDate	schema:deathDate	Text
foaf:member	foaf:member	Text
nombre alternativo	foaf:nick	Text
nombre alternativo 2	foaf:nick	Text
LOC	schema:hasExactExternalAuthority	URI
profesión	schema:occupationalCategory	Text

TABLA 5.5. Mapeo de metadatos de Autoridades Personas (Author-Authority)

5.3. Generación de URIs

Como plan para la creación de URIs se decidió crear una para cada individuo dentro de un archivo RDF, ya que, por tratarse de tesis propias de la Universidad de Chile, no es posible reutilizar una URI anterior o de otro sistema. Por otro lado, respecto a los académicos mencionados en los datos de muestra, solo tres cuentan con este dato de una fuente externa, es decir solo el 2 % del total de profesores mencionados en este proyecto. Se suma además a esta situación que los alumnos tesistas no cuentan con información biográfica dispuesta en algún sistema LOD. Finalmente se optó por incorporar URIs a fuentes externas cuando así fue necesario. Estos casos son principalmente cuando queremos establecer un vínculo o consular y extraer información para incorporar en el sistema local. Tal como menciona (Southwick, 2015) creamos URIs:

- por cosas que nos pertenecen (o de las que somos responsables) y que son únicas
- para agentes locales (personas u organismos corporativos) que no están controlados por la Biblioteca del Congreso u otras organizaciones de normalización

Por su parte la (W3C, 2014a) recomienda en “Los Principios de Diseño de la URI“ algunas normas como:

1. Utilice URIs HTTP para que puedan ser desreferenciados. La implicación para la generación de LOD es que para cada nueva URI creada, también es necesario crear un archivo RDF con la información sobre el recurso que identifica.
2. Proporcionar al menos una representación legible por máquina del recurso identificado por el URI.
3. Una estructura URI no debe contener nada que pueda cambiar.
4. La opacidad de las URI debe preservarse; es decir, no se deben separar las URI para inferir datos de ellas.

De acuerdo a lo anterior, y contando con la limitación de un subdominio pertinente, se optó simplemente por utilizar el nombre de servidor local, dejando a la URI con la siguiente estructura:

<servidor><namespace><clase>/<identificador único local>

Un ejemplo en este proyecto sería:

http://localhost/datos/recurso/rauch105000

5.4. Grafos RDF de ejemplo

A continuación se presenta una muestra de los archivos RDF que se han obtenido desde los cinco conjunto de datos obtenidos después del proceso de limpieza, reconciliación, y reestructuración a través del complemento para OpenRefine rdf-extension 0.8.

Se han obtenido los siguientes archivos RDF:

Para Recursos

```
<rdf:RDF
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:schema="http://schema.org/"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:foaf="http://xmlns.com/foaf/0.1/"
    xmlns:bibo="http://purl.org/ontology/bibo/"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:wgs84="http://www.w3.org/2003/01/geo/wgs84_pos#"
    xmlns:geonames="http://www.geonames.org/ontology#"
```

```
    xmlns:SKOS="http://www.w3.org/2004/02/SKOS/core#">

<rdf:Description rdf:about="http://localhost/datos/recurso/rauch105000\">
    <rdf:type rdf:resource="http://purl.org/ontology/bibo/Thesis"/>
    <dc:title>Caracterización y comparación anatómica de hojas de Peumo
    (Cryptocarya alba (Mol.) Looser) y Quillay (Quillaja saponaria Mol.)
    Sometidas a condiciones de riego permanente y de restricción hídrica</dc:title>
    <dc:date>2008</dc:date>
    <dc:language>Español</dc:language>
    <dc:type>Tesis</dc:type>
    <bibo:abstract>Peumo (Cryptocarya alba (Mol.) Looser) y quillay
    (Quillaja saponaria Mol.), son especies arbóreas nativas, que crecen
    en el área mediterránea de Chile, que se caracteriza por presentar
    veranos secos y calurosos. Por lo tanto las plantas están sometidas
    a restricción hídrica regularmente.
    Conocer las modificaciones anatómicas en estas especies como respuesta
    a la restricción hídrica, permitirá complementar estudios fisiológicos,
    cuyo fin es ayudar a comprender el efecto que tiene el ambiente sobre
    su desarrollo y productividad.
    El objetivo de este trabajo fue identificar algunas modificaciones
    anatómicas que se producen en las hojas completamente desarrolladas,
    de plantas jóvenes de peumo y quillay, causadas por la restricción
    hídrica. </bibo:abstract>
    <dc:description>
        Memoria para optar al Título Profesional de Ingeniero Forestal
    </dc:description>
    <dc:publisher>Universidad de Chile</dc:publisher>
    <bibo:uri>http://www.repositorio.uchile.cl/handle/2250/105000</bibo:uri>
    <dc:creator>Gotor Pedreros, Bárbara Francisca</dc:creator>
    <bibo:owner>Gotor Pedreros, Bárbara Francisca</bibo:owner>
    <dc:rights>Gotor Pedreros, Bárbara Francisca</dc:rights>
    <schema:author rdf:resource="http://localhost/datos/autoridad/person/au001"/>
    <dc:contributor>Donoso Calderón, Sergio</dc:contributor>
    <schema:contributor rdf:resource="http://localhost/datos/autoridad/person/au002"/>
    <dc:contributor>Peña Rojas, Karen</dc:contributor>
    <schema:contributor rdf:resource="http://localhost/datos/autoridad/person/au003"/>
    <dc:contributor>
```

```
        Facultad de Ciencias Forestales de la Universidad de Chile
    </dc:contributor>
    <schema:contributor rdf:resource=
    "http://es.dbpedia.org/resource/Facultad_de_Ciencias_Forestales
    _de_la_Universidad_de_Chile"/>
    <dc:contributor>Departamento de Silvicultura</dc:contributor>
    <dc:subject rdf:resource="http://localhost/datos/autoridad/subject/00001"/>
    <dc:subject rdf:resource="http://localhost/datos/autoridad/subject/00002"/>
    <dc:subject rdf:resource="http://localhost/datos/autoridad/subject/00003"/>
    <dc:subject rdf:resource="http://es.dbpedia.org/resource/Cryptocarya_alba"/>
    <dc:subject rdf:resource="http://es.dbpedia.org/resource/Quillaja_saponaria"/>
    <SKOS:notation>13115781</SKOS:notation>
</rdf:Description>
```

Ejemplo de SKOS-subject

```
<rdf:RDF\newline
    xmlns:schema="http://schema.org/"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:foaf="http://xmlns.com/foaf/0.1/"
    xmlns:bibo="http://purl.org/ontology/bibo/"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:SKOS="http://www.w3.org/2004/02/SKOS/core#">

<rdf:Description rdf:about="http://localhost/datos/autoridad/subject/00001">
    <rdf:type rdf:resource="http://www.w3.org/2004/02/SKOS/core#Concept"/>
    <SKOS:prefLabel>Riego</SKOS:prefLabel>
    <owl:sameAs rdf:resource="http://aims.fao.org/aos/agrovoc/c_3954"/>
    <bibo:uri rdf:resource=
    "http://artemide.art.uniroma2.it:8081/agrovoc/agrovoc/en/page/c_3954?clang=es"/>
    <SKOS:broader rdf:resource="http://localhost/datos/autoridad/subject/00088"/>
    <SKOS:altLabel>Irrigación</SKOS:altLabel>
    <SKOS:scopeNote>El riego consiste en aportar agua al suelo para que los vegetales
    tengan el suministro de agua que necesitan favoreciendo así su crecimiento.
    Se utiliza en la agricultura y en jardinería.</SKOS:scopeNote>
```

```
    <SKOS:narrower rdf:resource="http://localhost/datos/autoridad/subject/00027"/>
    <bibo:citedBy>Agrovoc</bibo:citedBy>
</rdf:Description>
```

Ejemplo de Places

```
<rdf:RDF
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:schema="http://schema.org/"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:dct="http://purl.org/dc/terms/"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:wgs84="http://www.w3.org/2003/01/geo/wgs84\_pos#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:geonames="http://www.geonames.org/ontology#"
    xmlns:SKOS="http://www.w3.org/2004/02/SKOS/core#">
<rdf:Description rdf:about="http://localhost/datos/place/topNCL000001">
    <rdf:type rdf:resource="http://www.geonames.org/ontology#Feature"/>
    <SKOS:prefLabel>Angol</SKOS:prefLabel>
    <wgs84:lat>-37.8043492</wgs84:lat>
    <wgs84:long>-72.7348068</wgs84:long>
    <owl:sameAs rdf:resource="http://es.dbpedia.org/resource/Angol"/>
    <owl:sameAs rdf:resource=
    "http://datos.bcn.cl/recurso/cl/division-politico-administrativa/2010/comuna/angol"/>
    <dct:isPartOf rdf:resource="http://localhost/datos/place/topNCLReg000001"/>
</rdf:Description>
```

Ejemplo de PlacesRegiones

```
<rdf:RDF
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:wgs84="http://www.w3.org/2003/01/geo/wgs84_pos#"
    xmlns:geonames="http://www.geonames.org/ontology#"
    xmlns:SKOS="http://www.w3.org/2004/02/SKOS/core#">
<rdf:Description rdf:about="http://localhost/datos/place/topNCLReg000001">
    <rdf:type rdf:resource="http://www.geonames.org/ontology#Feature"/>
```

```
    <SKOS:prefLabel>Región de la Araucanía</SKOS:prefLabel>
    <SKOS:altLabel>Región IX</SKOS:altLabel>
    <SKOS:altLabel>IX Región</SKOS:altLabel>
    <owl:sameAs rdf:resource="http://es.dbpedia.org/resource/Región_de_la_Araucanía"/>
</rdf:Description>
```

Ejemplo de Author-Authority

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
    xmlns:schema="http://schema.org/"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:foaf="http://xmlns.com/foaf/0.1/"
    xmlns:bio="http://purl.org/vocab/bio/0.1/"
    xmlns:madsrdf="http://www.loc.gov/mads/rdf/v1#"
    xmlns:bibo="http://purl.org/ontology/bibo/"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">

<rdf:Description rdf:about="http://localhost/datos/autoridad/person/au001">
    <dc:source rdf:resource="http://uchile.portafolio.academico"/>
    <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
    <foaf:name>Gotor Pedreros, Bárbara Francisca</foaf:name>
    <rdfs:label>Gotor Pedreros, Bárbara Francisca</rdfs:label>
    <foaf:birthday>s.f</foaf:birthday>
    <schema:deathDate>s.f</schema:deathDate>
    <bio:position>Tesista</bio:position>
    <foaf:nick>Bárbara Gotor</foaf:nick>
    <foaf:nick>Bárbara Gotor Pedreros</foaf:nick>
    <schema:occupationalCategory>Ingeniero Forestal</schema:occupationalCategory>
</rdf:Description>
```

5.5. Análisis de los grafos RDF

Una vez que se han obtenido los archivos RDF es necesario cargarlos al sistema Open Link Virtuoso. Los sitios más importantes en el desarrollo de los sistemas de Linked Open Data utilizan

este sistema, el cual es un sistema de base de datos que permite gestionar archivos rdf bajo el esquema de tripletas. Por lo tanto no es una base de datos relacional adaptada sino un sistema nativo para gestionar grafos rdf.

Se optó por Open Link Virtuoso pues existe la posibilidad de descargar una versión *open source* del sistema desde el sitio Github[1]. Permite la carga directa de archivos RDF, así como la interacción entre grafos, por otra parte es posible consultar los datos a través de un Sparql endpoint lo que facilita la integración con otras fuentes de información. Además al trabajar con proyectos escalables es posible llegar a 15.4 billones de tripletas de acuerdo a (WC3, 2015). Cuenta además con una interfaz gráfica vía Web del sistema la cual se puede visitar a través de http://localhost:8890 una vez instalado. Incluye ayuda en cada módulo del programa lo que facilita el uso de esta base de datos.

Junto con lo anterior es importante mencionar que proyectos como: DBpedia, la Biblioteca del Congreso de Chile, Biblioteca Nacional de España, entre muchos otros, utilizan OpenLink Virtuso como software de almacenamiento y consulta de tripletas. Por otro lado, la empresa que desarrolla el programa ofrece a través de su sitio Web la posibilidad de comprar el software para este tipo específico de base de datos RDF. Por último mencionar que este sistema es el que se utiliza en la Universidad de Chile para su proyecto de Datos Abiertos.

Los pasos a seguir para la instalación de virtuoso tal como se realizó para este proyecto, están indicados en el siguiente enlace: *http://carsten.io/virtuoso-os-on-mac-os/.*

Para saber más detalles de OpenLink Virtuso puede consultar el benchmark realizado por (W3C, 2016)

¿Cómo se han obtenido propiedades que permitan relacionar con otras fuentes de datos? Las fuentes de datos externas para este proyecto son: DBpedia, AGROVOC, Biblioteca del Congreso Nacional de Chile, Library of Congress. Esto es posible de observar en el grafo rdf **SKOS-subject** en el **object properties** *enlaceAgrovoc*, desde el cual es posible vincularse con AGROVOC. Por otro lado incorporar este tipo de herramienta de forma directa al sistema propuesto, permite contar con un alto grado de normalización de los temas tratados en las tesis:

object properties:

```
<owl:sameAs rdf:resource="http://aims.fao.org/aos/agrovoc/c_3954"/>
```

[1] `https://github.com/openlink/virtuoso-opensource`

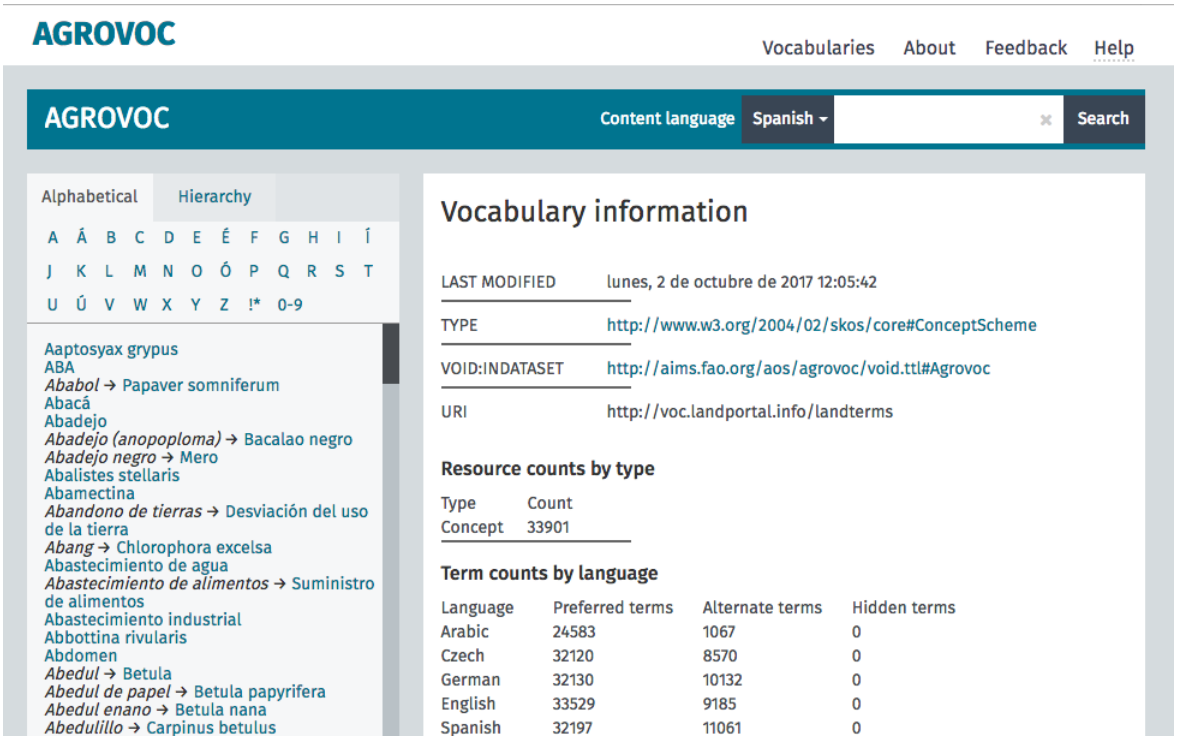


FIGURA 5.2. Sitio Web de AGROVOC de FAO.

El poder acceder al sitio a través del *object properties* es una puerta de acceso al resto de los datos que presenta este registro en el servidor externo. Para el presente ejemplo es posible acceder a los conceptos más específicos así como a los más genéricos, tal cual siguiendo la estructura de un tesauro figura 5.2 y figura 5.3.

A continuación es posible vincularse a DBpedia a través del **object properties** del grafo rdf **PlacesRegiones**:

```
<owl:sameAs rdf:resource="http://es.dbpedia.org/resource/Región_de_la_Araucanía"/>
```

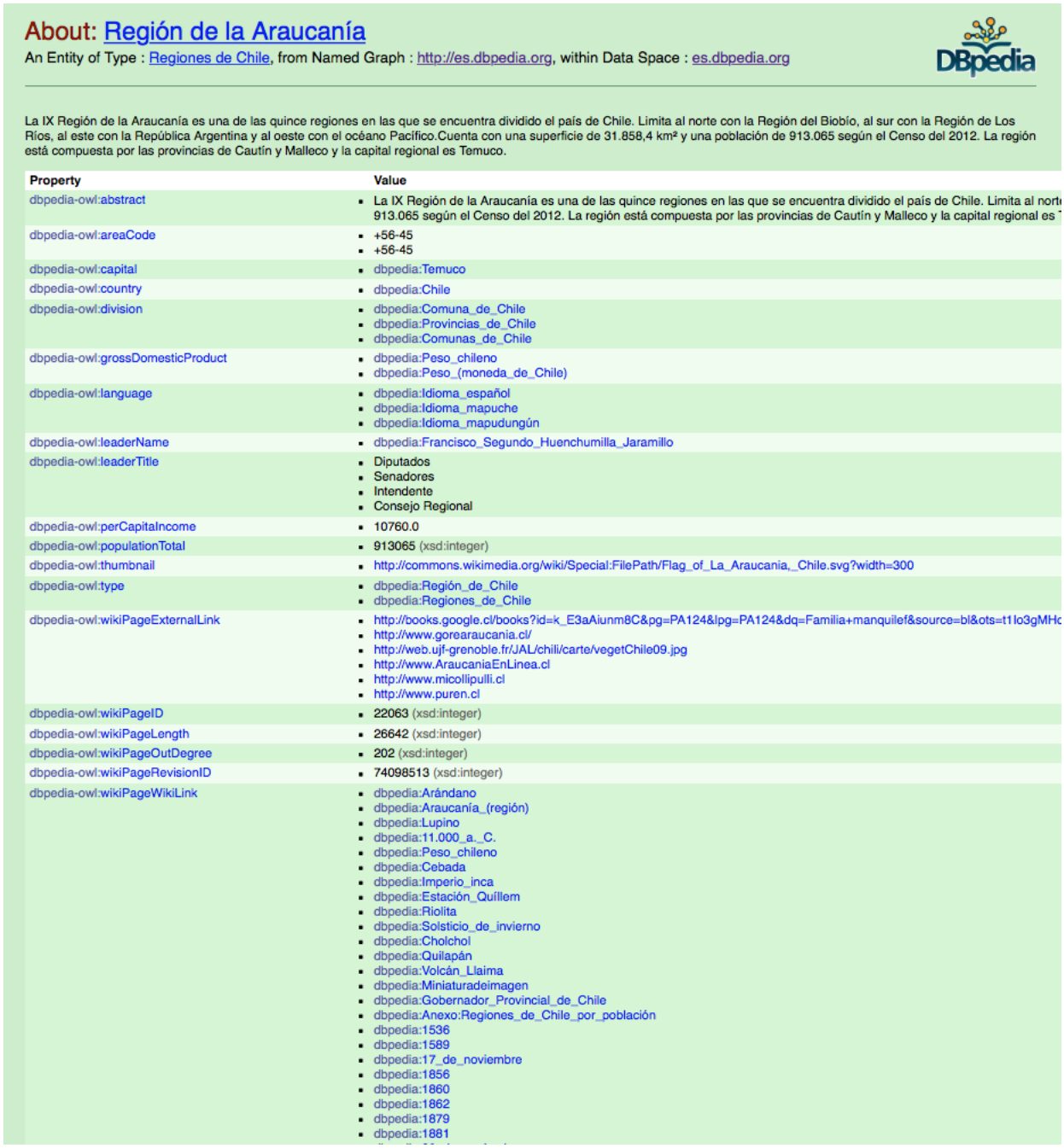
Igual que en el caso anterior el **object properties** se comporta como un punto de acceso al resto de la información que compone el registro de la Región de la Araucanía en DBpedia. Por lo que sería posible integrar información a la interfaz de usuario, como: cuál es la capital de esta región con el *data properties* dbpedia-owl:capital o cual es la población del lugar a través del **data properties dbpedia-owl:populationTotal**, figura 5.4.

http://aims.fao.org/aos/agrovoc/c_3954
irrigation

Property	Value
rdf:type	skos:Concept
skos:inScheme	http://aims.fao.org/aos/agrovoc
skos:broader	http://aims.fao.org/aos/agrovoc/c_330834 http://aims.fao.org/aos/agrovoc/c_330834
skos:narrower	http://aims.fao.org/aos/agrovoc/c_15970 http://aims.fao.org/aos/agrovoc/c_25323 http://aims.fao.org/aos/agrovoc/c_25352 http://aims.fao.org/aos/agrovoc/c_25353 http://aims.fao.org/aos/agrovoc/c_25355 http://aims.fao.org/aos/agrovoc/c_35247 http://aims.fao.org/aos/agrovoc/c_8332 http://aims.fao.org/aos/agrovoc/c_25354
skos:exactMatch	http://lod.nal.usda.gov/nalt/24709 http://stitch.cs.vu.nl/vocabularies/rameau/ark:/12148/cb119321398 http://zbw.eu/stw/descriptor/10490-0 http://eurovoc.europa.eu/2680 http://cat.aii.caas.cn/concept/c_16025 http://lod.gesis.org/thesoz/concept/10039159 http://aims.fao.org/aos/asfa/c_6741 http://d-nb.info/gnd/4006306-9 http://linkeddata.ge.imati.cnr.it:2020/resource/EARTh/52290 http://cat.aii.caas.cn/concept/c_15964 http://www.eionet.europa.eu/gemet/concept/4505
skos:closeMatch	http://dbpedia.org/resource/Irrigation http://purl.org/bncf/tid/5008
	http://cat.aii.caas.cn/concept/c_15992 http://cat.aii.caas.cn/concept/c_15990 http://cat.aii.caas.cn/concept/c_15985

prefLabel	altLabel	Lang
آبیاری		fa
Bewaesserung	Bewässerung	de
灌漑		ja
관개		ko
závlaha		sk
Irrigation		fr
Irrigazione		it
závlaha		cs
灌溉		zh
Riego	Irrigación	es
sulama		tr
орошение		ru
зрошування	зрошення	uk
irrigation		en
Nawadnianie		pl
सिंचाई		hi
ري		ar
การชลประทาน		th
నీటి పారుదల		te
öntözés		hu
ຊົນລະປະທານ		lo
Irrigação		pt

FIGURA 5.3. Sitio Web datos abiertos enlazados de AGROVOC de FAO.



About: Región de la Araucanía

An Entity of Type : Regiones de Chile, from Named Graph : http://es.dbpedia.org, within Data Space : es.dbpedia.org

La IX Región de la Araucanía es una de las quince regiones en las que se encuentra dividido el país de Chile. Limita al norte con la Región del Biobío, al sur con la Región de Los Ríos, al este con la República Argentina y al oeste con el océano Pacífico.Cuenta con una superficie de 31.858,4 km² y una población de 913.065 según el Censo del 2012. La región está compuesta por las provincias de Cautín y Malleco y la capital regional es Temuco.

Property	Value
dbpedia-owl:abstract	La IX Región de la Araucanía es una de las quince regiones en las que se encuentra dividido el país de Chile. Limita al nort… 913.065 según el Censo del 2012. La región está compuesta por las provincias de Cautín y Malleco y la capital regional es …
dbpedia-owl:areaCode	+56-45 +56-45
dbpedia-owl:capital	dbpedia:Temuco
dbpedia-owl:country	dbpedia:Chile
dbpedia-owl:division	dbpedia:Comuna_de_Chile dbpedia:Provincias_de_Chile dbpedia:Comunas_de_Chile
dbpedia-owl:grossDomesticProduct	dbpedia:Peso_chileno dbpedia:Peso_(moneda_de_Chile)
dbpedia-owl:language	dbpedia:Idioma_español dbpedia:Idioma_mapuche dbpedia:Idioma_mapudungún
dbpedia-owl:leaderName	dbpedia:Francisco_Segundo_Huenchumilla_Jaramillo
dbpedia-owl:leaderTitle	Diputados Senadores Intendente Consejo Regional
dbpedia-owl:perCapitaIncome	10760.0
dbpedia-owl:populationTotal	913065 (xsd:integer)
dbpedia-owl:thumbnail	http://commons.wikimedia.org/wiki/Special:FilePath/Flag_of_La_Araucania,_Chile.svg?width=300
dbpedia-owl:type	dbpedia:Región_de_Chile dbpedia:Regiones_de_Chile
dbpedia-owl:wikiPageExternalLink	http://books.google.cl/books?id=k_E3aAiunm8C&pg=PA124&lpg=PA124&dq=Familia+manquilef&source=bl&ots=t1lo3gMHc… http://www.gorearaucania.cl/ http://web.ujf-grenoble.fr/JAL/chili/carte/vegetChile09.jpg http://www.AraucaniaEnLinea.cl http://www.micollipulli.cl http://www.puren.cl
dbpedia-owl:wikiPageID	22063 (xsd:integer)
dbpedia-owl:wikiPageLength	26642 (xsd:integer)
dbpedia-owl:wikiPageOutDegree	202 (xsd:integer)
dbpedia-owl:wikiPageRevisionID	74098513 (xsd:integer)
dbpedia-owl:wikiPageWikiLink	dbpedia:Arándano dbpedia:Araucanía_(región) dbpedia:Lupino dbpedia:11.000_a._C. dbpedia:Peso_chileno dbpedia:Cebada dbpedia:Imperio_inca dbpedia:Estación_Quillem dbpedia:Riolita dbpedia:Solsticio_de_invierno dbpedia:Cholchol dbpedia:Quilapán dbpedia:Volcán_Llaima dbpedia:Miniaturadeimagen dbpedia:Gobernador_Provincial_de_Chile dbpedia:Anexo:Regiones_de_Chile_por_población dbpedia:1536 dbpedia:1589 dbpedia:17_de_noviembre dbpedia:1856 dbpedia:1860 dbpedia:1862 dbpedia:1879 dbpedia:1881

FIGURA 5.4. Sitio Web datos abiertos enlazados DBpedia.

5.6. Frontend del sistema para tesis

El frontend se carga en primera instancia las últimas tesis ingresadas al sistema, presenta una ordenación por fecha de publicación tal como muestra la figura 5.5.

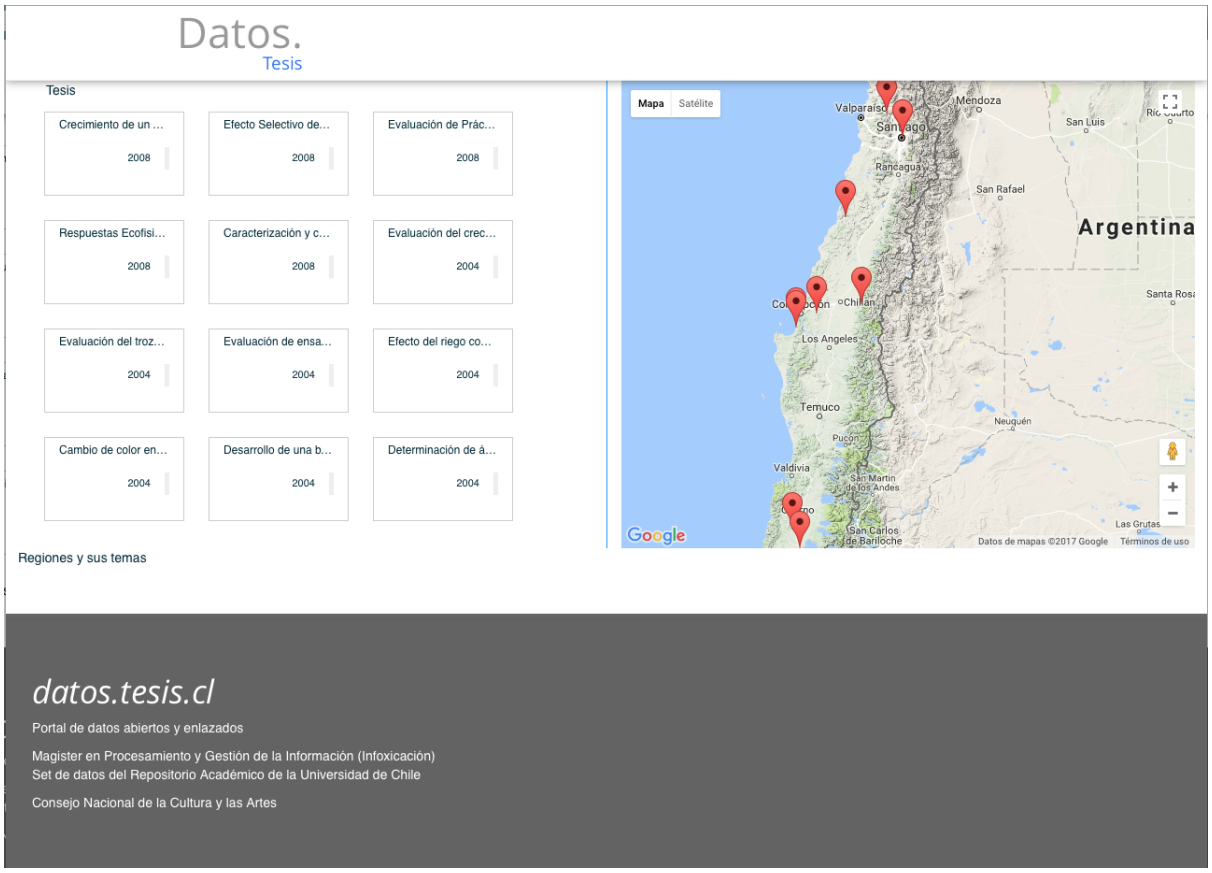


FIGURA 5.5. Sitio Web datos abiertos enlazados del proyecto.

FIGURA 5.6. Sitio Web datos abiertos enlazados del proyecto.

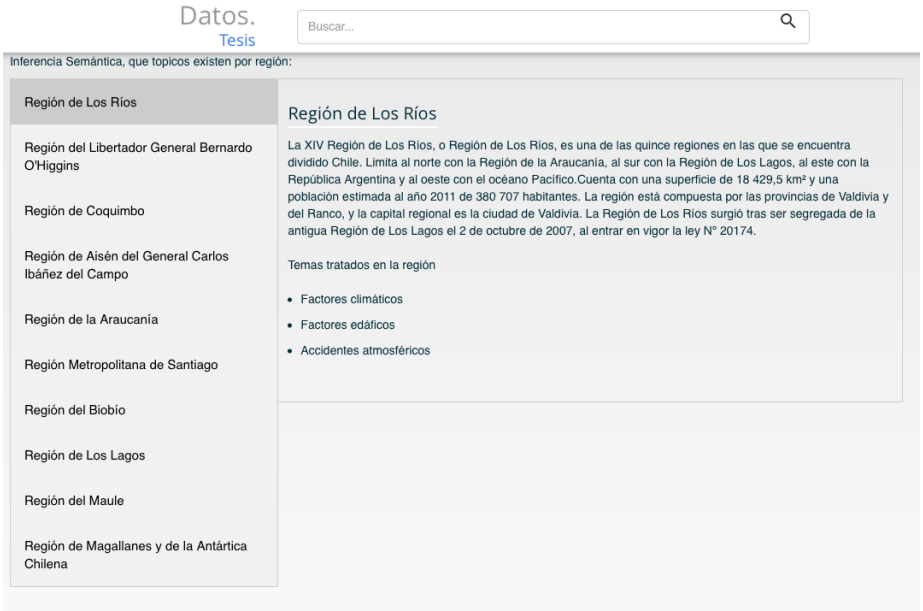


FIGURA 5.7. Ejercicio de inferencia semántica.

5.7. Linked Open Data inserto en el actual Repositorio Académico

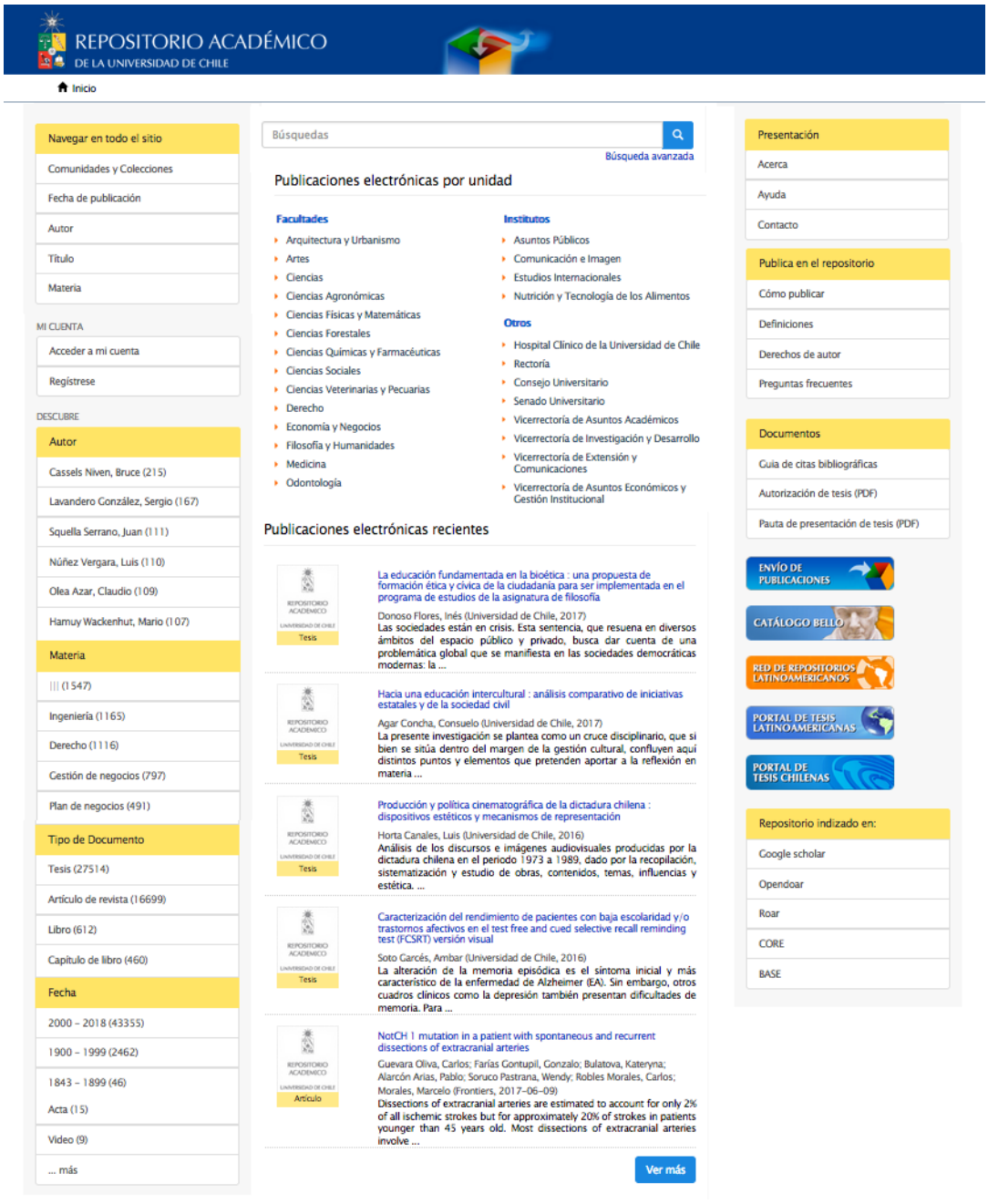


FIGURA 5.8. Actual repositorio académico de la Universidad de Chile.

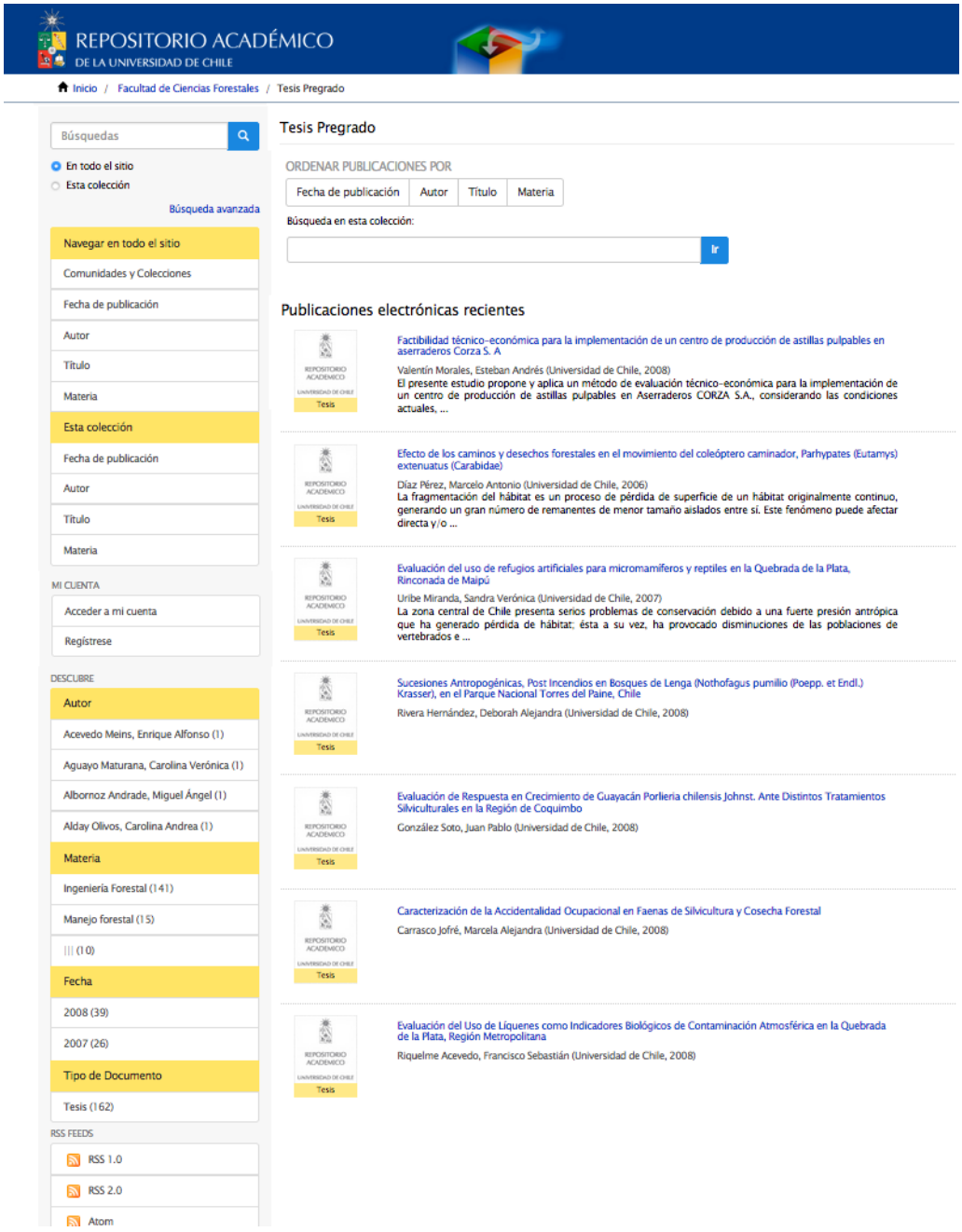


FIGURA 5.9. Despliegue actual de tesis de pregrado de la Facultad de Ciencias Forestales.

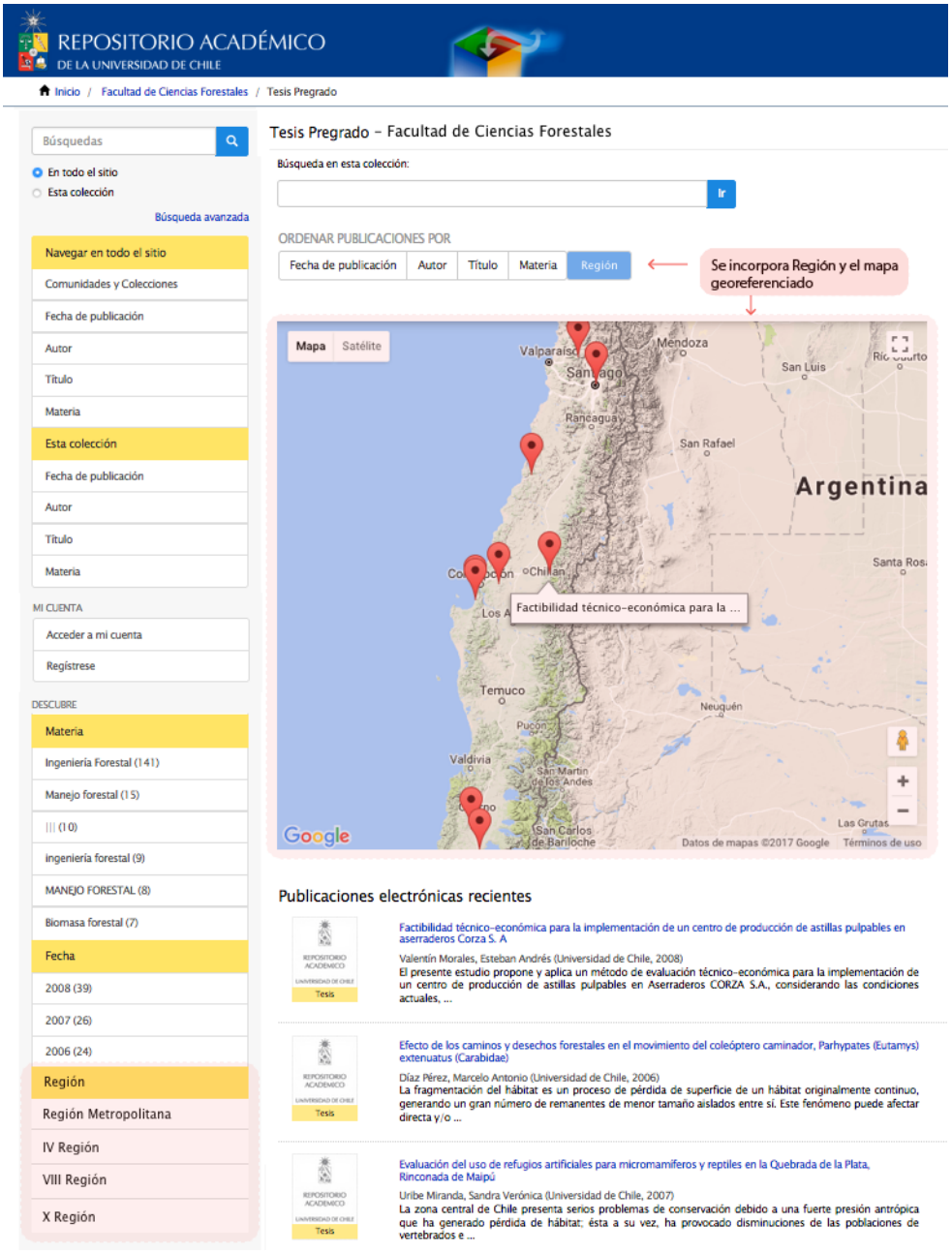


FIGURA 5.10. Interfaz que incorpora el potencial semántico en la visualización, región en las facetas y el mapa georeferenciado.

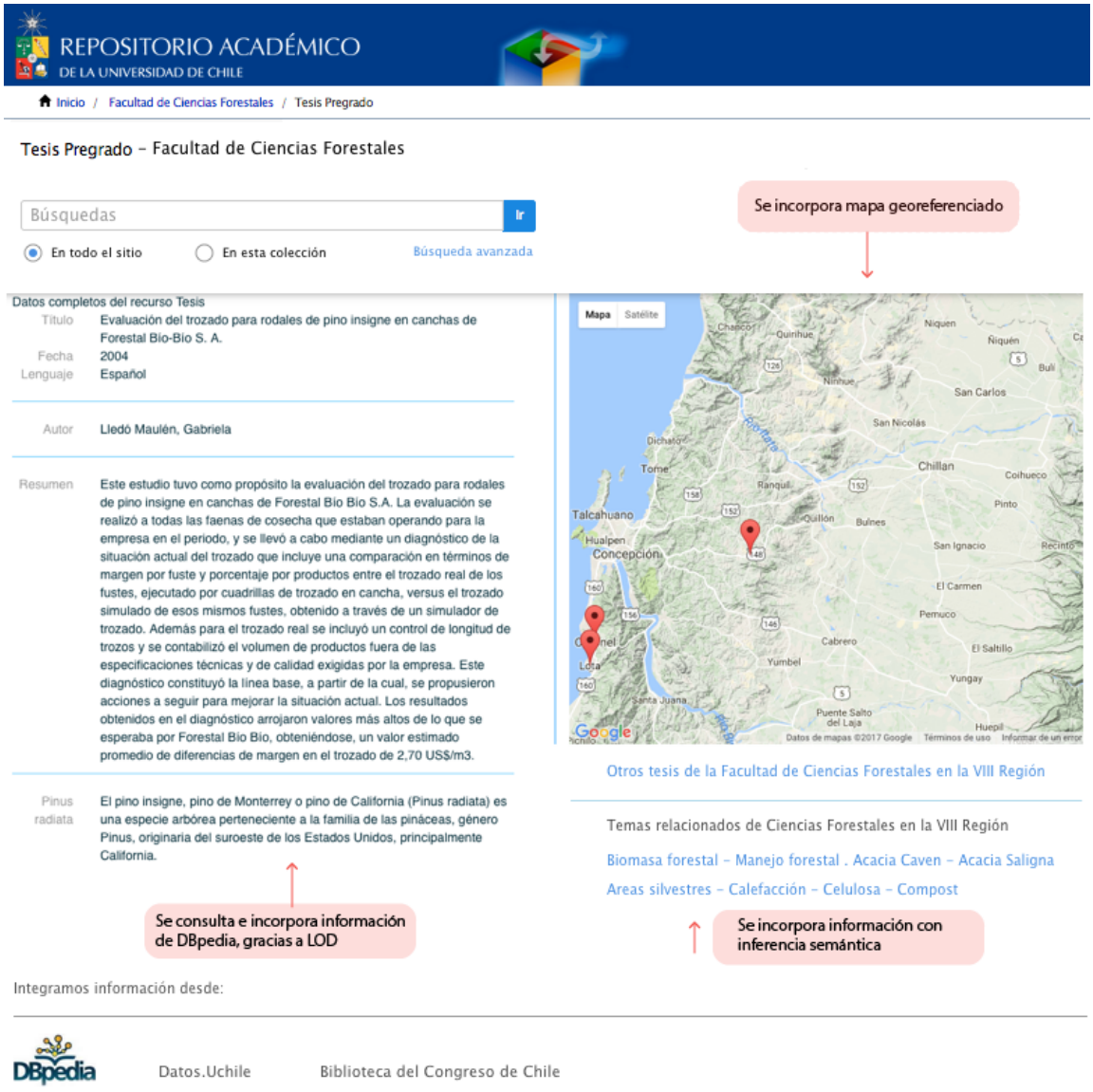


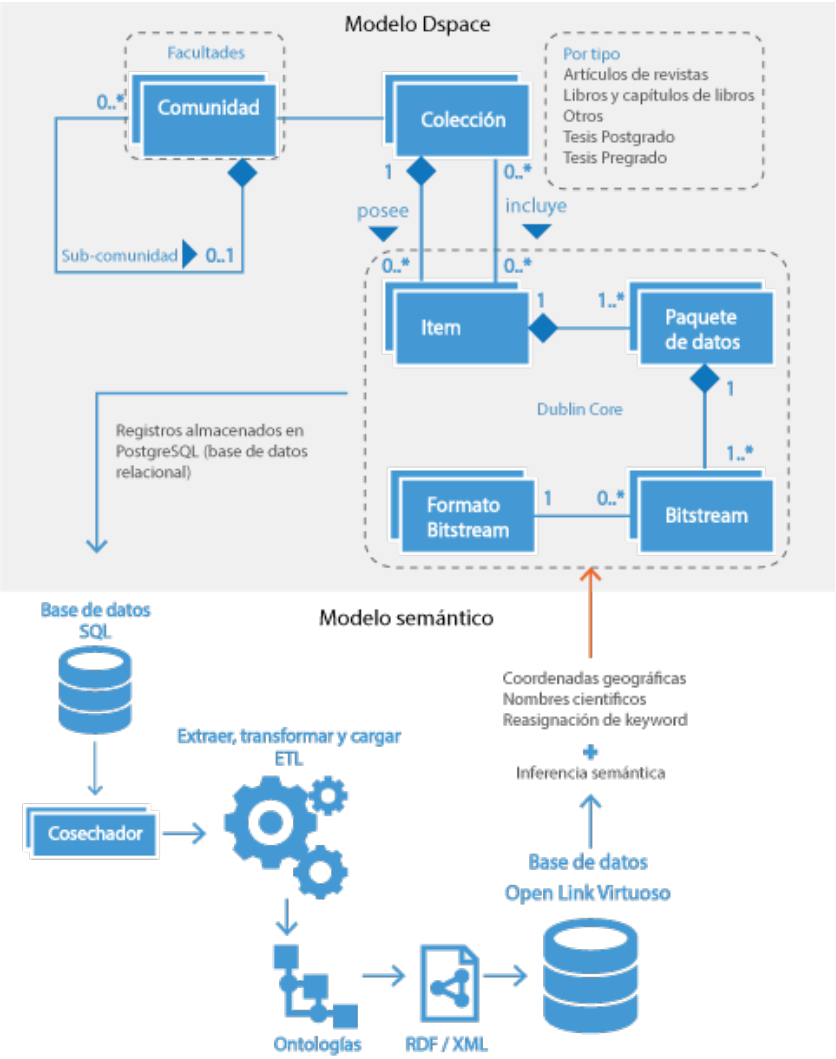
FIGURA 5.11. Visualización de un registro que incorpora el nombre científico en tiempo real desde DBpedia como a su vez un mapa señalando la ubicación geográfica en donde se realizó el estudio; incluye también la posibilidad de acceder a otros documentos de la misma región con otras temáticas del área forestal.

5.8. Modelo de datos del Repositorio de la U. de Chile incorporando el modelo semántico

El modelo actual del repositorio académico se basa en la estructura que ofrece Dspace a través de Comunidades, Sub-comunidades y Colecciones, lo que se traduce en:

Comunidades	Facultades o departamentos
Subcomunidades	no se utiliza
Colecciones	Artículos de revistas Libros y capítulos de libros Otros Tesis Postgrado Tesis Pregrado
Item	Bitstream (archivos, pdf, word, excel, etc.)

TABLA 5.6. Uso de Dspace en Repositorio Académico.



En esta figura se puede apreciar cómo el modelo semántico se podría incorporar al actual modelo, enriqueciendo la información a través de la incorporación de DBpedia y con datos previamente normalizados. Para esto es necesario:

- La extracción a través de un cosechador cómo MarcEdit.
- Luego un ETL como OpenRefine el cual permite, normalizar y transformar los registros de Dspace incorporando las ontologías para generar los archivos RDF.
- Carga de archivos RDF a base de datos de grafos OpenLink Virtuoso.
- Programar la lectura de datos desde los archivos RDF en 'Virtuso' a Dspace, esto se traduce en complementar la base de datos relacional con el sistema semántico.

6. CONCLUSIONES

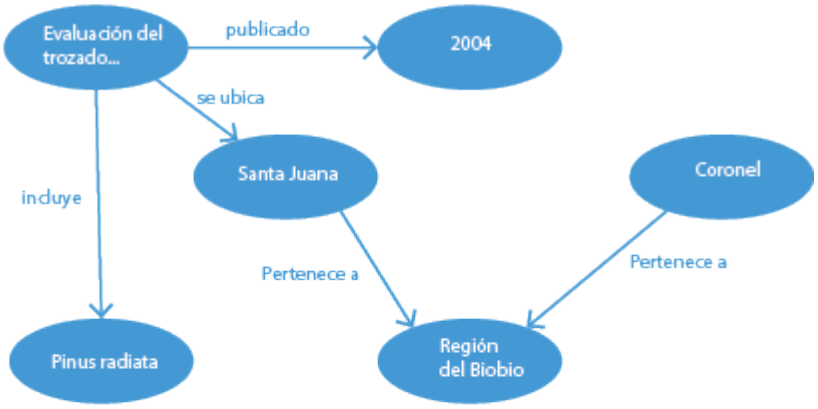
Al finalizar este trabajo de investigación, donde se abordó una secuencia de pasos, necesarios para modelar una muestra de registros bibliográficos de tesis de la Facultad de Ciencias Forestales de la Universidad de Chile a un sistema RDF, es posible obtener las siguientes conclusiones:

El repositorio académico de la Universidad de Chile,es la fuente de información patrimonial que recoge y pone a disposición de la comunidad de usuarios las tesis de la institución a través del sistema Dspace, utilizando la estructura de metadatos Dublin Core para la descripción de la información. Para trabajar los registros bibliográficos de estas tesis, primero fue necesario seleccionar una muestra acotada de registros ya que el tiempo involucrado para limpiar, normalizar y transformarlos a RDF fue considerable. Una muestra acotada permite probar y reacomodar metodologías optimizando el trabajo con el fin de alcanzar el objetivo planteado en esta investigación, expresado finalmente en la aplicación de tecnología semántica.

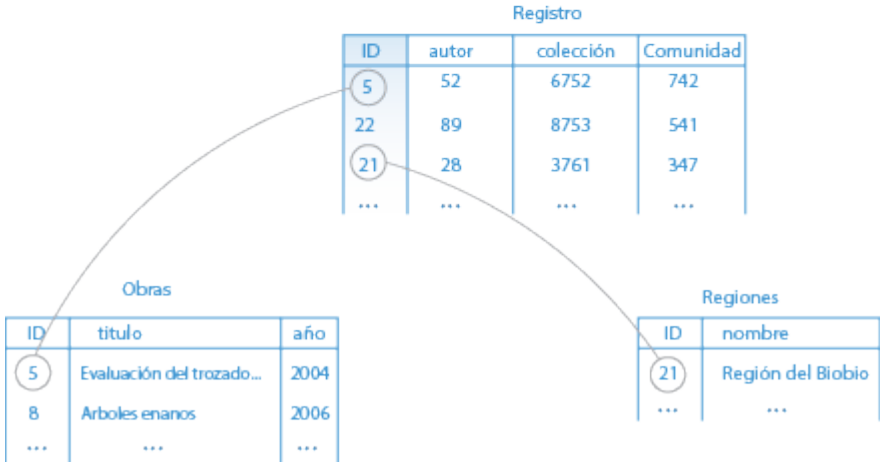
Para contar con puntos de comparación con el ejercicio realizado, fue necesario buscar y reconocer experiencias de aplicación de linked open data en otros sistemas bibliográficos como: Biblioteca Nacional de Francia, Biblioteca Nacional de España, Biblioteca del Congreso en Chile entre otras, esto permitió establecer puntos en común o diferencias en cómo se muestran los datos de un registro. A su vez, un aspecto relevante fue encontrar a DBpedia como la fuente de información externa más enlazada por parte de los distintos proyectos investigados.

Se pudo comprobar que un sistema de base de datos semántico guarda significativas diferencias con una base de datos relacional, como Dspace, lo cual se explica en la siguiente figura:

Base de datos semántica



Base de datos relacional



Un sistema semántico establece relaciones con significado entre los datos, todas las semánticas se hacen explícitas por la tripleta en sí. Al hacerlo, ya no es necesario que el esquema interprete los datos. Dentro del mundo de las bases de datos y XML, solo los datos que se ajustan a las reglas definidas en el esquema, pueden existir y codificarse en la base de datos o archivo XML. Con RDF, solo se hacen declaraciones sobre hechos que se conoce, pero estas declaraciones pueden interactuar con declaraciones hechas fuera de su sistema de información, como lo mencionado con DBpedia. Este modelo de datos, además permite que los datos heterogéneos se conecten e interactúen. En el caso de un modelo relacional, se registran los datos en un formato tabular donde cada registro contiene una clave única por tabla, comúnmente se le denomina ID, este elemento se comparte con otros registros que la pueden utilizar para referirse a una relación. De esta forma se vinculan las distintas tablas. Estos sistemas se organizan en entidades, atributos y vínculos.

Para poder crear los archivos RDF, fue necesario la incorporación de las siguientes ontologías: SKOS este se utiliza para la estructuración de vocabulario controlado principalmente de contenidos, bibo ontología utilizada para describir información bibliográfica, Schema permite describir y calificar los tipos de documentos, FOAF permite describir la relación entre personas. BIO se complementa con FOAF, ya que es posible incluir información biográfica de un individuo, entre otras posibilidades. GeoNames, es una ontología asociada a nombres geográficos, WGS84 se complementa con la anterior, la que permite incluir por ejemplo puntos geográficos exactos a través de latitud,longitud. dc (es el sistema de metadatos Dublin Core) junto con dcterms que se refiere a Dublin Core calificado.

Por otro lado, las ontologías incorporadas con el fin de obtener un modelo de datos RDF que incluya una descripción de información más completa de lo que ofrece actualmente el repositorio, para esto fue fundamental el estudio de los modelos expuestos por Europeana y la Biblioteca Nacional Británica, instituciones que han hecho pública la información de sus proyectos de linked open data. Esto permitió definir la subdivisión de registros que componían la muestra inicial en cinco diferentes grafos RDF: Skos-subject para el control de materias, Places para lugares geográficos específicos, PlacesRegiones para lugares geográficos generales, Author-Authority para los nombres de personas, y Recursos para la descripción general de una obra, lo anterior permitió mejorar el control de la información, como también definir los puntos de enlace con fuentes externas a través del object properties correspondiente.

Para una correcta obtención de uno o más archivos RDF es posible utilizar OpenRefine, con su complemento rdf-extension para modelar los registros y traducirlos a RDF de manera automatizada. Es importante destacar que para lograr este objetivo se deben incorporar las distintas ontologías mencionadas en el punto anterior. Luego de contar con los archivos RDF, estos se agregan a la base de datos que gestiona los Grafos, como Openlink Virtuoso, en este etapa ya se podrían realizar consultas y búsquedas de información sobre el sistema a través de Sparql EndPoint que provee 'Virtuoso'.

Al concluir este ejercicio de titulación, es posible dimensionar la importancia de cada uno de los pasos para llegar al resultado final, la motivación de este estudio siempre fue el generar una mejora en calidad de la información disponible por la Universidad de Chile, a través de su repositorio académico. Al mismo tiempo, sensibilizar a la comunidad profesional bibliotecaria a incorporar estas nuevas herramientas en su trabajo, las cuales ya están presentes en diversos servicios web que diariamente se consumen, y que las tecnologías de información nos llevan a usar. Se espera contribuir a incorporar tecnología semántica por las bibliotecas en Chile y avanzar en brindar información de calidad a los diversos usuarios de todo el mundo.

BIBLIOGRAFIA

Auer, S., Lehmann, J., y Hellmann, S. (2009). Linkedgeodata: Adding a spatial dimension to the web of data. En A. Bernstein et al. (Eds.), *The semantic web - iswc 2009: 8th international semantic web conference, iswc 2009, chantilly, va, usa, october 25-29, 2009. proceedings* (pp. 731–746). Berlin, Heidelberg: Springer Berlin Heidelberg. Descargado de https://doi.org/10.1007/978-3-642-04930-9_46 doi: 10.1007/978-3-642-04930-9_46

Baca, M. (2016). *Introduction to metadata*. Getty Publications.

Breitbach, W. (2016, apr). *Web-scale discovery: Utopian dream or dystopian nightmare (or maybe something in between)?* California Academic and Research Libraries.

Bui, Y., y Park, J.-r. (2006). An assessment of metadata quality: A case study of the national science digital library metadata repository. En *Proceedings of the annual conference of cais/actes du congrès annuel de l'acsi.*

Castillo Guerrero, R. (2008). *infoxicacion.* Descargado 2017-10-27, de http://www.infoxicacion.cl/vocabulario-controlado-soporte-de-la-estructuracion-de-contenidos/

Chen, Y.-N. (2017). A review of practices for transforming library legacy records into linked open data. En E. Garoufallou, S. Virkus, R. Siatri, y D. Koutsomiha (Eds.), *Metadata and semantic research: 11th international conference, mtsr 2017, tallinn, estonia, november 28 – december 1, 2017, proceedings* (pp. 123–133). Cham: Springer International Publishing. Descargado de https://doi.org/10.1007/978-3-319-70863-8_12 doi: 10.1007/978-3-319-70863-8_12

Cole, T. W., Han, M.-J., Weathers, W. F., y Joyner, E. (2013). Library marc records into linked open data: Challenges and opportunities. *Journal of Library Metadata*, *13*(2-3), 163-196. Descargado de https://doi.org/10.1080/19386389.2013.826074 doi: 10.1080/19386389.2013.826074

De Boer, V., Wielemaker, J., Van Gent, J., Hildebrand, M., Isaac, A., Van Ossenbruggen, J., y Schreiber, G. (2012). Supporting linked data production for cultural heritage institutes: the amsterdam museum case study. *The Semantic Web: Research and Applications*, 733–747.

Di Noia, T., Ragone, A., Maurino, A., Mongiello, M., Marzocca, M. P., Cultrera, G., y Bruno, M. P. (2016). Linking data in digital libraries: the case of puglia digital library. En *Whise@ eswc* (pp. 27–38).

Donohue, T. (2017). *Metadata and bitstream format registries.* Descargado 2017-11-05, de `https://wiki.duraspace.org/display/DSDOC6x/Metadata+and+Bitstream+Format+Registries`

Duraspace. (2016, sep). *Dspace metadata rdf mapping vocabulary.* `http://digital-repositories.org/ontologies/dspace-metadata-mapping/`. ((Accessed on 10/25/2017))

FAO. (2014). *Agrovoc tesauro multilingüe de agricultura | agricultural information management standards (aims).* Descargado 2017-10-23, de `http://aims.fao.org/es/agrovoc`

Goddard, L., y Byrne, G. (2010). The strongest link: Libraries and linked data. *D-Lib magazine*, *16*(11/12).

Google. (2012, may). *Official google blog: Introducing the knowledge graph: things, not strings.* `https://googleblog.blogspot.cl/2012/05/introducing-knowledge-graph-things-not.html`. ((Accessed on 10/25/2017))

Gómez-Castaño, J., Barrueco Cruz, J. M., París-Folch, M.-L., Aguilar-Lorente, E., y Martínez-Galindo, F. J. (2015). *Los repositorios institucionales de las universidades públicas valencianas: situación actual y retos para el futuro.* Descargado 2017-10-18, de `http://helvia.uco.es/handle/10396/12625`

Heath, T., y Bizer, C. (2011). *Linked data: Evolving the web into a global data space*. Morgan & Claypool.

Konstantinou, N., Houssos, N., y Manta, A. (2014). Exposing bibliographic information as linked open data using standards-based mappings: methodology and results. *Procedia-Social and Behavioral Sciences*, *147*, 260–267.

Konstantinou, N., Spanos, D.-E., Houssos, N., y Mitrou, N. (2014). Exposing scholarly information as linked open data: Rdfizing dspace contents. *The Electronic Library*, *32*(6), 834–851.

Lubas, R., Jackson, A., y Schneider, I. (2013). *The metadata manual: A practical workbook*. Elsevier Science.

NISO. (2010). *Niso standards - national information standards organization.* Descargado 2017-10-27, de `http://www.niso.org/kst/reports/standards?step=2&gid=None&project_key=7cc9b583cb5a62e8c15d3099e0bb46bbae9cf38a`

OAI. (2017). *Open archives initiative.* Descargado 2017-10-19, de `https://www.openarchives.org/`

OpenDOAR. (2017). *Opendoar chart - usage of open access repository software - worldwide.* Descargado 2017-11-27, de `http://www.opendoar.org`

Openrefine. (2017). Descargado 2017-10-19, de `http://openrefine.org/`

Pastor-Sánchez, J.-A., y Saorín, T. (2015). Web semántica. informe de situación 2014. *Anuario Think EPI*.

Pomerantz, J. (2015). *Metadata*. The MIT Press.

Radio, E., y Hanrath, S. (2016). Measuring the impact and effectiveness of transitioning to a linked data vocabulary. *Journal of Library Metadata*, *16*(2), 80–94.

Rdf 1.1 primer. (2014). Descargado 2017-10-22, de `https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/`

Reese, T. (2013). *Marcedit development.* Descargado 2017-10-19, de `http://marcedit.reeset.net/`

Rodríguez-Bravo, B., Simoes, M.-d.-G., Vieira-de Freitas, M.-C., y Frías, J.-A. (2017, mayo-junio). *Descubrimiento de información científica: ¿todavía misión y visión de la biblioteca académica? | rodríguez-bravo | el profesional de la información.* https://recyt.fecyt.es/index.php/EPI/article/view/epi.2017.may.13. ((Accessed on 10/18/2017))

Scholar, G. (2017). *Google scholar help.* Descargado 2017-12-24, de https://scholar.google.com/intl/en/scholar/inclusion.html#indexing

Segaran, T., Evans, C., y Taylor, J. (2009). *Programming the semantic web: Build flexible applications with graph data*. O'Reilly Media.

Sikos, L. (2015). *Mastering structured data on the semantic web: From html5 microdata to linked open data*. Apress.

Southwick, S. B. (2015, mar). A guide for transforming digital collections metadata into linked data using open source technologies. *Journal of Library Metadata*, *15*(1), 1-35.

Steele, T., y Sump-crethar, N. (2016). Metadata for electronic theses and dissertations: A survey of institutional repositories. *Journal of Library Metadata*, *16*(1), 53–68.

Tay, A. (2016a). *5 thoughts on open access, institutional and subject repositories.* Descargado 2017-10-18, de http://musingsaboutlibrarianship.blogspot.cl/2016/10/5-thoughts-on-open-access-institutional.html

Tay, A. (2016b). *Are institutional repositories a dead end?* Descargado 2017-10-12, de http://musingsaboutlibrarianship.blogspot.cl/2016/08/are-institutional-repositories-failing.html

Tay, A. (2016c). *Library discovery and the open access challenge - take 2.* Descargado 2017-10-18, de http://musingsaboutlibrarianship.blogspot.cl/2016/12/library-discovery-and-open-access.html?utm_source=feedburner&utm_medium=email&utm_campaign=Feed:+MusingsAboutLibrarianship+(Musings+about+librarianship)

Uniform resource identifier - wikipedia. (s.f.). https://en.wikipedia.org/wiki/Uniform_Resource_Identifier. ((Accessed on 10/12/2017))

Universidad de Chile, S. (2016). *Acerca.* Descargado 2017-10-15, de http://repositorio.uchile.cl/page/acerca

Universidad de Chile, S. (2017). *Red de repositorios latinoamericanos - universidad de chile.* Descargado 2017-11-27, de http://repositorioslatinoamericanos.uchile.cl/

Van de Velde, E. (2016). *Let ir rip.* Descargado 2017-10-15, de http://scitechsociety.blogspot.cl/2016/07/let-ir-rip.html

Van Hooland, S., y Verborgh, R. (2014). *Linked data for libraries, archives and museums: How to clean, link and publish your metadata*. Neal-Schuman. Descargado de https://books.google.cl/books?id=NjamoAEACAAJ

Verborgh, R., y De Wilde, M. (2013). *Using openrefine:*. Packt Publishing.

W3C. (2014a). *Best practices for publishing linked data.* Descargado 2017-10-27, de https://www.w3.org/TR/ld-bp/

W3C. (2014b). *Rdf primer.* Descargado 2017-10-28, de https://www.w3.org/TR/2004/REC-rdf-primer-20040210/

W3C. (2016). *Rdf store benchmarking - w3c wiki.* Descargado 2017-12-27, de https://www.w3.org/wiki/RdfStoreBenchmarking

WC3. (2015). *Large triple stores - w3c wiki.* Descargado 2017-12-27, de https://www.w3.org/wiki/LargeTripleStores

Wood, D., Zaidman, M., Ruth, L., y Hausenblas, M. (2014). *Linked data*. Manning Publications Co.

Yoshimura, K. S. (2016). Analysis of international linked data survey for implementers. *D-Lib Magazine*, *22*(7), 6.

ANEXO A. CONCEPTOS ASOCIADOS A DATOS ABIERTOS ENLAZADOS

AGROVOC: se forma de la unión de las palabras agricultura y vocabulario. Es un vocabulario controlado que abarca todos los ámbitos de interés de la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO[1]), entre ellos la alimentación, la nutrición, la agricultura, la pesca, las ciencias forestales y el medio ambiente. Lo publica la FAO y una comunidad de expertos se encarga de su edición. El vocabulario consiste en más de 32.000 conceptos con hasta 40.000 términos en 23 lenguas. AGROVOC está disponible como un esquema RDF[2]/SKOS[3]-XL y está publicado como un conjunto de datos enlazados, lo cual lo hace más atractivo de usar en este proyecto.

Datos enlazados: datos y conjuntos de datos compartidos en la Web abierta y que contienen enlaces a otros datos utilizando tecnologías Web estándar. Definición según (Pomerantz, 2015).

Dspace[4]: es el software para las organizaciones académicas, sin fines de lucro y comerciales que crean repositorios digitales abiertos. Es completamente personalizable para adaptarse a las necesidades de cualquier organización. DSpace conserva y permite el acceso fácil y abierto a todo tipo de contenido digital, incluyendo texto, imágenes, imágenes en movimiento, mpegs y conjuntos de datos. Y con una comunidad cada vez mayor de desarrolladores, comprometidos a expandir y mejorar continuamente el software, cada instalación de DSpace se beneficia de la siguiente. Este sistema utiliza Java para su matriz y PostgreSQL como gestor de bases de datos.

Dublin Core: un conjunto de elementos desarrollado para ser el conjunto básico necesario para describir cualquier recurso en línea. Existen dos versiones una mínima con quince metadatos[5] y otra extendida Iniciativa de Metadatos Dublin Core[6] con cincuenta y cinco metadatos.

Marcedit[7]: programa computacional que permite extraer, limpiar, analizar registros bibliográficos desde distintos sistemas computacionales, pudiendo realizar cosechas desde sistemas MARC, OAI-PMH, OpenRefine, u otros. Se encuentran disponibles para distintas plataformas, como: Windows, OSX, Linux.

[1] http://www.fao.org/home/es/
[2] https://es.wikipedia.org/wiki/Resource_Description_Framework
[3] https://es.wikipedia.org/wiki/Simple_Knowledge_Organization_System
[4] http://www.dspace.org/introducing
[5] http://dublincore.org/documents/dces/
[6] http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=terms
[7] http://marcedit.reeset.net/

Metadatos: los metadatos están a nuestro alrededor, todo el tiempo. En la era moderna de la electrónica ubicua, cámaras fotográficas, archivos pdf o cualquier otro, computadores personales, casi todos los dispositivos que usa se basan en metadatos o lo generan, o ambos. Como nos dice (Pomerantz, 2015): “pero cuando los metadatos están haciendo su trabajo bien, simplemente se desvanece en el fondo, desapercibido y casi invisible”. Metadatos se definen genéricamente como dato acerca de los datos.

OpenRefine: OpenRefine[8] (anteriormente Google Refine) de acuerdo al sitio que lo pone a disposición, lo define como una poderosa herramienta para trabajar con datos desordenados: limpiarlo; transformándolo a distintos formatos; y ampliarlo con servicios Web y datos externos. Para conocer más de esta aplicación puede utilizar el libro (Verborgh y De Wilde, 2013).

OWL: el W3C Web Ontology Language (OWL) es un lenguaje Web Semántico diseñado para representar un conocimiento rico y complejo sobre cosas, grupos de cosas y relaciones entre cosas. OWL es un lenguaje basado en la lógica computacional tal que el conocimiento expresado en OWL puede ser explotado por programas informáticos, por ejemplo, para verificar la consistencia de ese conocimiento o para hacer explícito el conocimiento implícito. Los documentos de OWL, conocidos como ontologías, pueden ser publicados en la World Wide Web y pueden referirse o ser referidos de otras ontologías de OWL. OWL es parte de la pila de tecnología Web semántica del W3C, que incluye RDF, RDFS, sparql, etc. (Extraído desde el sitio https://www.w3.org/OWL/)

Resource Description Framework (RDF): de acuerdo a (Pomerantz, 2015): es un modelo de datos, en otras palabras, es un marco, una estructura lógica según la cual los datos están organizados. ¿Un marco para qué? Para describir los recursos. ¿Qué recursos? Cualquier recurso en general, aunque generalmente se utiliza RDF para describir recursos en la Web". En resumen, RDF es un modelo genérico de datos para hacer declaraciones descriptivas sobre entidades. Por otro lado (Sikos, 2015) nos dice: “un RDF se puede utilizar para crear una descripción interpretable por la máquina sobre cualquier tipo de recurso Web, ya que los archivos RDF pueden ampliarse con un número arbitrario de vocabularios externos. De hecho, RDF y otros estándares básicos de la Web Semántica como RDFS y OWL tienen sus propios vocabularios, que usualmente se combinan entre sí y se amplían con otros vocabularios para describir objetos y sus propiedades. Es importante tener en cuenta, sin

[8] `http://openrefine.org/`

embargo, que RDF es mucho más que un vocabulario, ya que es un lenguaje de modelado de datos semánticos completo.

Sparql endpoint: un servicio que acepta consultas sparql y devuelve respuestas a ellas como conjuntos de resultados sparql, esto se puede utilizar en forma remota o local.

Sparql y Lenguaje de consultas RDF: un estándar de lenguaje de consulta para datos RDF en la Web Semántica; Análogo al lenguaje de consulta estructurado (SQL) para las bases de datos relacionales.

Triplestore: en general, las bases de datos RDF son triplestores. Un triplestore es un sistema que tiene alguna forma de almacenamiento persistente de datos RDF y le permite ejecutar consultas sparql contra esos datos. Según (Wood et al., 2014) "Muchos de estos sistemas se basan en la reutilización y adaptación de técnicas bien establecidas a partir de bases de datos relacionales. Un enfoque básico sería crear una sola tabla de tres columnas donde la primera columna contenga el sujeto, la segunda el predicado y la tercera columna el objeto. La tabla básica de tres columnas puede optimizarse, pero aún no existen mejores prácticas. Las técnicas de optimización incluye la indexación, las aplicaciones de varias funciones hash y la reestructuración del enfoque básico de la tabla."

Tripleta o Triple: a las sentencias RDF, también llamadas triples, pueden conectarse entre sí para formar grafos, compuestos de sujeto-predicado-objetos, en algunos casos se les nombran con URIs. El modelo de datos RDF define lo que cada componente de un triple puede ser, como un URI o un literal, cuando hablamos del modelo **sujeto-predicado-objeto** También define otros conceptos clave tales como cómo restringir literales con tipos de datos o en qué idioma (humano) están escritos. Algunos componentes del modelo de datos, denominados con URIs, pueden ser recolectados en clases para que puedan ser más fácilmente descubiertos, buscados o consultados. De acuerdo a (Wood et al., 2014) define triple como:"Una declaración RDF, que consta de dos cosas (un sujeto y un objeto) y una relación entre ellos (un verbo, o predicado). Este objeto-predicado-objeto triple forma el más pequeño posible grafo RDF (aunque la mayoría de los grafos RDF consisten en muchas de tales declaraciones)".

Sujeto	Predicado	Objeto
<http://misitioweb.org/datos_climaticos>	rdfs:label	Observaciones ambientales

TABLA A.1. Estructura de tripleta

Ontología: Según (Wood et al., 2014) se define como un modelo formal que permite representar el conocimiento para un dominio específico. Por su parte (Pomerantz, 2015) dice que una ontología es como un tesauro, un conjunto finito de términos, organizados como una jerarquía que se puede utilizar para proporcionar un valor para un elemento. Además, esto incluye un conjunto de reglas para la acción, a menudo en forma de algoritmos de software.

Open Archive Initiative[9] Reutilización e intercambio de objetos (OAI-ORE): define estándares para la descripción y el intercambio de agregaciones de recursos Web. Estas agregaciones, a veces llamadas objetos digitales compuestos, pueden combinar recursos distribuidos con múltiples tipos de medios incluyendo texto, imágenes, datos y video. El objetivo de estas normas es exponer el contenido enriquecido de estas agregaciones a aplicaciones que admiten creación, depósito, intercambio, visualización, reutilización y preservación. Aunque un caso de uso motivador para el trabajo es la naturaleza cambiante de la erudición y la comunicación académica, y la necesidad de la ciber-infraestructura para apoyar esa beca, la intención del esfuerzo es desarrollar estándares que generalizan a través de toda la información basada en la Web, Redes de "Web 2.0".

Uniform Resource Identifier: Según (*Uniform Resource Identifier - Wikipedia*, s.f.) es “Es una cadena de caracteres utilizada para identificar un recurso”.”Tal identificación permite la interacción con las representaciones del recurso sobre una red, típicamente la World Wide Web, usando protocolos específicos. Los esquemas que especifican una sintaxis concreta y protocolos asociados definen cada URI”. Por otro lado (Baca, 2016) lo define brevemente como: cadena corta que identifica de forma única un recurso como un documento HTML, una imagen, un archivo descargable o un servicio.

XML[10]: La definición que entrega la W3C es; Extensible Markup Language (XML) es un formato de texto simple, muy flexible derivado de SGML (ISO 8879). Originalmente diseñado para satisfacer los desafíos de la publicación electrónica a gran escala, XML también está desempeñando un papel cada vez más importante en el intercambio de una amplia variedad de datos en la Web y en otros lugares.

[9] `http://www.openarchives.org/ore/`
[10] `https://www.w3.org/XML/`

ANEXO B. DUBLIN CORE CUALIFICADO UTILIZADO POR DSPACE

TABLA B.1. Dublin Core Calificado (Donohue, 2017)

Elemento	Calificado	Nota de alcance
contributor		A person, organization, or service responsible for the content of the resource. Catch-all for unspecified contributors.
contributor	advisor	Use primarily for thesis advisor.
contributor	author(*)	Author(s) of the work (used by default)
contributor		
contributor	illustrator	
contributor	other	
coverage	spatial	Spatial characteristics of content.
coverage	temporal	Temporal characteristics of content.
creator		May be used as an alternative to çontributor.author"
date		Use qualified form if possible.
date	accessioned(*)	Date DSpace takes possession of item.
date		Date or date range item became available to the public.
date	copyright	Date of copyright.
date	created	Date of creation or manufacture of intellectual content if different from date.issued.
date	issued1	Date of publication or distribution.
date	submitted	Recommend for theses/dissertations.
identifier		
identifier	citation(**)	Human-readable, standard bibliographic citation of non-DSpace format of this item
identifier	govdoc(**)	A government document number
identifier	isbn(**)	International Standard Book Number
identifier	issn(**)	International Standard Serial Number
identifier	sici	Serial Item and Contribution Identifier
identifier	ismn(**)	International Standard Music Number

TABLA B.1. Dublin Core Calificado (Donohue, 2017)

Elemento	Calificado	Nota de alcance
identifier	other(**)	A known identifier type common to a local collection.
identifier	uri(*)	Uniform Resource Identifier
description(*)		Catch-all for any description not defined by qualifiers.
description	abstract(*)	Abstract or summary.
description	provenance(*)	The history of custody of the item since its creation, including any changes successive custodians made to it.
description	sponsorship(**)	Information about sponsoring agencies, individuals, or contractual arrangements for the item.
description	statementofresponsibility	To preserve statement of responsibility from MARC records.
description	tableofcontents	A table of contents for a given item.
description	uri	Uniform Resource Identifier pointing to description of this item.
format2		Catch-all for any format information not defined by qualifiers.
format	extent(**)	Size or duration.
format	medium(**)	Physical medium.
format	mimetype(**)	Registered MIME type identifiers.
language		Catch-all for non-ISO forms of the language of the item, accommodating harvested values.
language	iso(**)	Current ISO standard for language of intellectual content, including country codes (e.g. .en_US").
publisher(**)		Entity responsible for publication, distribution, or imprint.
relation		Catch-all for references to other related items.
relation	isformatof	References additional physical form.
relation	ispartof	References physically or logically containing item.
relation1	ispartofseries	Series name and number within that series, if available.
relation	haspart	References physically or logically contained item.
relation	isversionof	References earlier version.
relation	hasversion	References later version.
relation	isbasedon	References source.

TABLA B.1. Dublin Core Calificado (Donohue, 2017)

Elemento	Calificado	Nota de alcance
relation	isreferencedby	Pointed to by referenced resource.
relation	requires	Referenced resource is required to support function, delivery, or coherence of item.
relation	replaces	References preceeding item.
relation	isreplacedby	References succeeding item.
relation	uri	References Uniform Resource Identifier for related item
rights		Terms governing use and reproduction.
rights	uri	References terms governing use and reproduction.
source		Do not use; only for harvested metadata.
source	uri	Do not use; only for harvested metadata.
subject(**)		Uncontrolled index term.
subject	classification	Catch-all for value from local classification system. Global classification systems will receive specific qualifier
subject	ddc	Dewey Decimal Classification Number
subject	lcc	Library of Congress Classification Number
subject	lcsh	Library of Congress Subject Headings
subject	mesh	MEdical Subject Headings
subject	other	Local controlled vocabulary; global vocabularies will receive specific qualifier.
title(*)		Title statement/title proper.
title	alternative(**)	Varying (or substitute) form of title proper appearing in item, e.g. abbreviation or translation
type(*)		Nature or genre of content.

(*) Utilizado por varias áreas funcionales de DSpace. NO SE QUITAN SIN INVESTIGAR LAS CONSECUENCIA

(**) Este campo se incluye en la interfaz de usuario predeterminada de DSpace Submission. Si elimina este campo de su registro, se romperá el formulario de envió predeterminado de DSpace.

ANEXO C. MODELO DE DATOS EUROPEANA

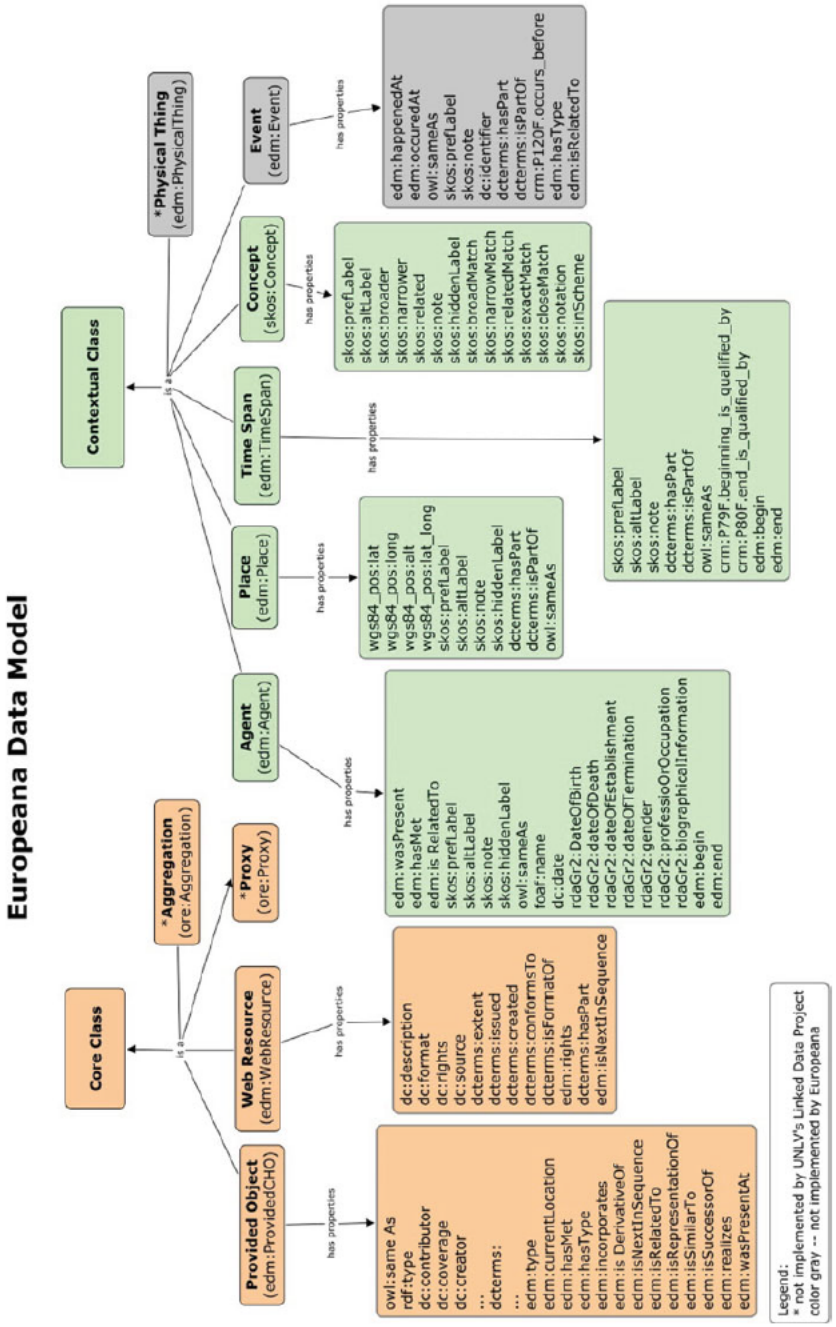


FIGURA C.1. Modelo de datos Europeana.

ANEXO D. MODELO DE DATOS DE BRITISH LIBRARY

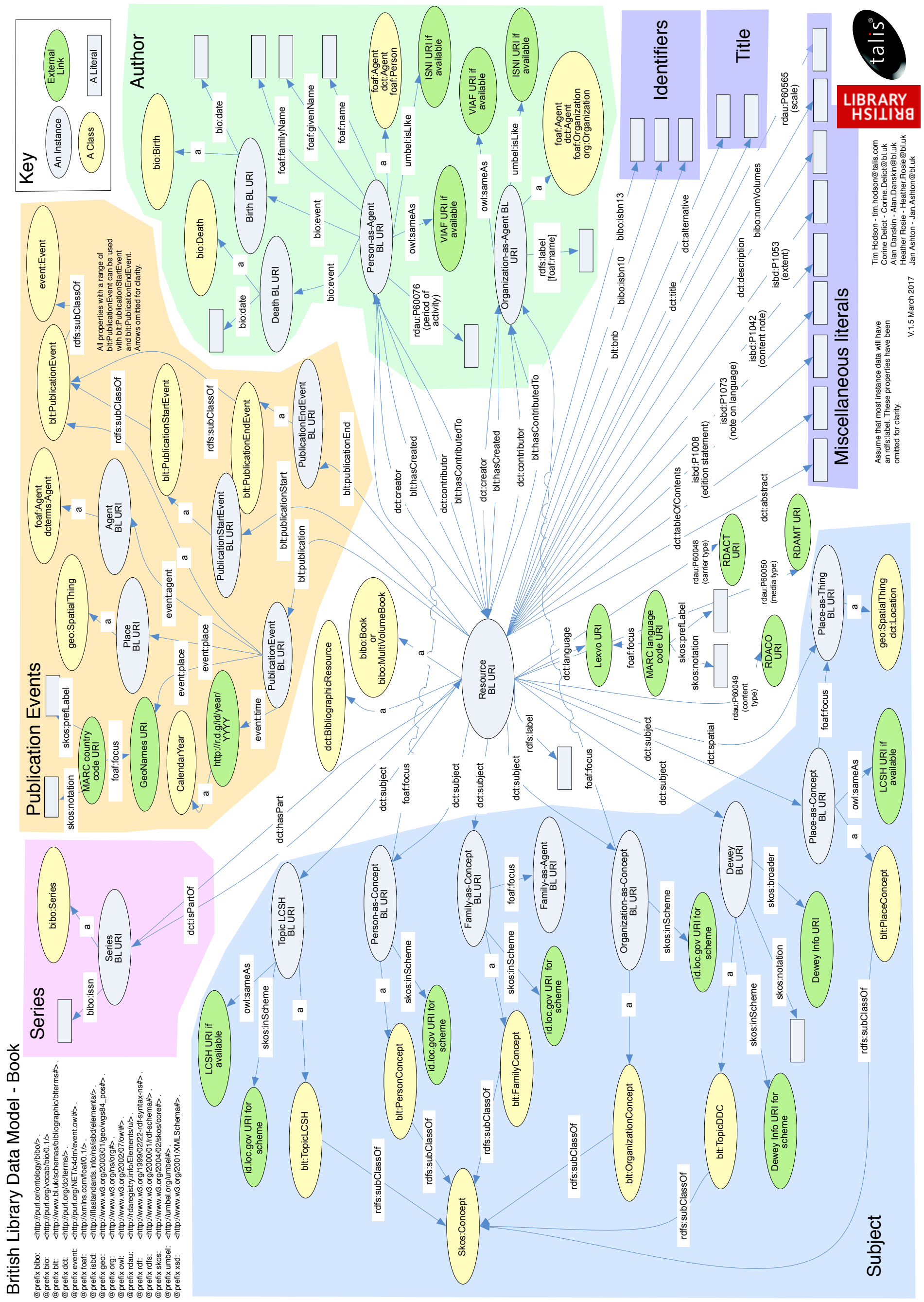


FIGURA D.1. Modelo de datos Biblioteca Británica.