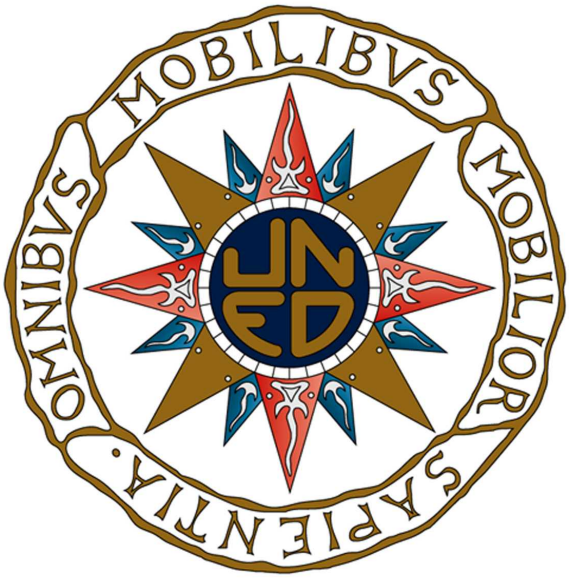


E.T.S. DE INGENIERÍA INFORMÁTICA

DEPARTAMENTO DE INTELIGENCIA ARTIFICIAL

Máster Universitario en Inteligencia Artificial Avanzada: Fundamentos, Métodos y Aplicaciones

Diseño de un *datastore* sobre datos académicos de la UNED y su enriquecimiento vía Minería de Textos desde el corpus de Guías Docentes

Carlos Arancón del Valle

Directores:
José Luis Fernández Vindel
Víctor Fresno Fernández

2017

Agradecimientos

Quiero dedicar este trabajo a mi familia y en especial a mi madre Asun. Sólo recordarte que eres la mujer más fuerte que conozco y que si lo superamos hace siete años, lo volveremos a superar sin duda. ¡Ánimo! Te quiero mucho...

A mi pareja Cris. Gracias por quererme tal como soy, ayudarme, comprenderme y apoyarme en todo, por soñar a mi lado, y por dedicar tu presente para que podamos escribir un futuro juntos. ¡Te quiero!

A mis padres Santiago y Asunción, de quienes he aprendido los valores que hoy me guían. Gracias por darlo todo para que no nos faltara nada, por estar siempre ahí, por vuestra dedicación y ayuda incondicional. Os quiero mucho.

A mis hermanas Cristina y Estela, con quienes he crecido. Gracias por compartir conmigo tantos momentos (tanto buenos, como no tan buenos) a lo largo de toda una vida. Aunque no lo diga muy a menudo, os quiero un montón.

Gracias a mis amigos por formar parte de mi vida y de mi entorno. Da gusto desconectar del mundo a vuestro lado.

Gracias a mis tutores, por vuestra predisposición y ayuda. Sin vosotros este trabajo no hubiera sido posible.

A José Luis Fernández Vindel, por tu paciencia infinita hasta encontrar un tema que me encajara y por la gran aportación de ideas, tanto a este trabajo, como a todas las líneas que se quedaron abiertas. Has sido un gran tutor desde que comenzamos con el modelado de procesos con Wings hasta que acabamos con el presente trabajo, al que has dado un gran valor con tu enfoque.

A Víctor Fresno Fernández, por todo el tiempo que has invertido en ayudarme a resolver mis dudas y corregir mis errores (que no ha sido poco). He aprendido mucho de las decenas de correos que tuviste que padecer sobre las funciones de extracción terminológica, su modificación, su evaluación... He disfrutado mucho con este proyecto.

Por último, a Omar Khalil Gómez y a Aitor Díaz Medina por vuestra implicación en la creación de una interfaz de navegación terminológica que explotara algunos de los resultados de este trabajo, aportándole un gran valor añadido. También quiero dar las gracias a los docentes que habéis dedicado unos minutos de vuestro tiempo para colaborar en la elaboración del *Gold Standard* y a todos aquellos que de una manera u otra habéis colaborado haciendo posible este trabajo.

Muchísimas gracias a todos.

Índice general

Índice de figuras

Índice de tablas

Resumen

Este Trabajo de Fin de Máster se concibe bajo el desarrollo en tres etapas diferenciadas, con alcances distintos:

1. El diseño e implementación de un *datastore* RDF funcional, poblado con datos académicos de la UNED y sobre el que se puedan ejecutar consultas complejas.

2. La recopilación automática de recursos documentales referenciados en ese *datastore*, generando un corpus sobre el cual se realizará la extracción de información de interés desde estas fuentes no estructuradas, y su posterior estructuración e integración de vuelta en el *datastore* inicial.

3. Una reflexión sobre los vocabularios usados en esta experiencia y su alineamiento con ontologías más generales aplicables a los agentes, recursos y procesos académicos en la universidad.

En la sección 3.1 se presenta la primera etapa con una prueba de concepto, muy guiada por las vistas públicas de los datos UNED a nuestro alcance, cuyas clases y relaciones se trasladan al modelo RDF. Se configura así un datastore exhaustivamente poblado: con todos los datos sobre estructura, personal y oferta académica de la UNED, con sus interrelaciones. La institución no dispone, hasta la fecha, de un punto público de consulta similar.

A partir de este desarrollo, en la sección 3.2 se plantea el objetivo básico de investigación de este trabajo: el diseño y evaluación de funciones de extracción terminológica (para fines específicos) en el corpus de Guías de Estudio referenciadas desde el *datastore*. El *datastore* de partida contiene referencias a recursos que pueden ser automáticamente recuperados y analizados mediante técnicas de Minería de Datos, con las que llevar a cabo una extracción terminológica que desemboque en nueva información con la que enriquecer el *datastore*.

Finalmente, el enriquecimiento del *datastore* con estos términos requería una mínima ampliación de los vocabularios iniciales, que además conviene alinear con ontologías externas para facilitar su uso. Esta revisión se produce en la sección 3.3, donde se configura una tercera etapa en la que se inicia una discusión integrada de ontologías relevantes para describir los procesos universitarios. Por acotación temporal se asume como un objetivo secundario, así como una descripción ampliada de trabajos futuros.

Capítulo 1

Introducción

1.1. Motivación y Contexto

Hoy en día con el crecimiento de la información publicada en la Web por numerosas instituciones y usuarios, surge la necesidad de una administración eficiente de ésta, que facilite el acceso, reutilización y explotación de los datos publicados. Tim Berners-Lee presenta el concepto de Web Semántica como una extensión de la Web actual, en la cual se da un significado bien definido a la información estableciendo enlaces entre los datos, del mismo modo que en la Web actual se establecen enlaces entre las páginas (hiperenlaces) [6]. A raíz de la noción de Web Semántica y con el objetivo de conectar, publicar y acceder a los datos procedentes de distintas fuentes, surge el concepto de Datos Enlazados (Linked Data) con una serie de principios que constituyen las buenas prácticas a seguir para la correcta generación de este tipo de datos [5]. Si además los datos son de acceso público (Datos Enlazados Abiertos), las ventajas que trae consigo este modelo crecen considerablemente.

Muchas son las instituciones que han decidido publicar sus datos en abierto de manera que puedan ser fácilmente accesibles por cualquier usuario, ya sea mediante consultas directas o mediante aplicaciones y sistemas de explotación de esta información. Ejemplos de estas instituciones son: el Banco Mundial[1] que publica, entre otros, datos abiertos sobre su actividad; WikiPedia[2] a través de DBpedia[3]; City-Go-Round[4] que comparte datos sobre el transporte en grandes ciudades; Freebase[5] con multitud de información obtenida de diversas fuentes; Wikidata[6] que recopila datos estructurados para dar soporte, entre otros, a Wikipedia y a Wikimedia Commons. Otros ejemplos a nivel nacional son la Agencia Estatal de Meteorología[7], el Instituto Geográfico Nacional[8] e incluso el Gobierno de España[9].

Pero tan importante y necesario como la publicación adecuada de los datos, es una explotación eficiente de éstos. Tanto es así, que los sistemas inteligentes con mayor impacto en los últimos años se apoyan en la gestión masiva de datos, bien porque en su entrenamiento se requieren, bien porque su respuesta se construye, en cada momento, a partir de este conocimiento masivo accesible.

Más allá del nivel de publicidad de los datos, cada empresa, institución o movimiento social genera su propia traza de datos digitales. Estas organizaciones pueden llevar a cabo una gestión y explotación dirigida de datos accesibles (internos o externos), donde se pueden observar varias

[1] http://datos.bancomundial.org/
[2] https://es.wikipedia.org/wiki/Wikipedia:Portada
[3] http://wiki.dbpedia.org/
[4] http://www.citygoround.org/
[5] http://www.freebase.com
[6] https://www.wikidata.org
[7] http://www.aemet.es
[8] http://www.ign.es
[9] http://datos.gob.es/

aproximaciones:

- Por un lado, las que ofrecen las Tecnologías Semánticas, donde se reformulan los datos explícitos de la organización (usualmente exportados de sus bases de datos relacionales) y se modelan sus jerarquías conceptuales y las reglas de inferencia aplicables. Todo ello desde la herencia de décadas de trabajo en IA simbólica, adaptada a los estándares de gestión de la Web Semántica y con el objetivo de permitir un crecimiento flexible, una interconexión pública y un enriquecimiento con datos externos.

- Por otro lado, toda la proyección actualizada de la IA conexionista y otras técnicas de análisis, con el objetivo genérico de identificar patrones de interés sobre datos desestructurados. Entre esas técnicas se encuentran todas las aplicables a la Minería de Textos. Y también los procesos de minería de datos y aprendizaje automático sobre trazas muy masivas (logs de servidores, conexiones, etc) que se han popularizado bajo la etiqueta imprecisa de Big Data.

Un dominio donde la utilización de Datos Enlazados Abiertos adquiere gran potencial, es el educativo, y de manera específica el universitario. Varias universidades se han sumado a participar del beneficio que proporciona esta iniciativa.

En la UNED, sin embargo, no existe hasta la fecha un *datastore* RDF[10] público más allá del proyecto de Datos Enlazados UNEDATA[11], que se creó con la intención de llevar a cabo una explotación eficaz de toda la información disponible sobre recursos dedicados a publicaciones académicas y comunidades de investigación. Sin embargo, dicho proyecto se encuentra escasamente poblado y no ofrece un punto de consulta SPARQL donde usuarios o agentes puedan acceder sencillamente a esta información.

Por otro lado, la UNED facilita de forma abierta listados y vistas públicas de algunos de sus datos en el Portal de Transparencia[12], y otros listados y vistas más estructuradas que sólo son accesibles por los miembros de la institución (tras su autenticación) para fines colectivos diversos. Una de las carencias funcionales más evidentes de los portales citados se da en torno a las facilidades de consulta abierta establecidas para personas y agentes automáticos.

Este trabajo se empieza a diseñar desde una hipótesis de viabilidad técnica: la capacidad que tenían los directores y el autor de recopilar datos públicos de la UNED (sobre organización, profesorado y oferta académica), de agregarlos, filtrarlos y producir un *datastore* RDF básico con todos estos datos, en un primer vocabulario propio.

Desde este *datastore* se plantean varias líneas de trabajo posibles. Todas ellas, en este marco de Datos Enlazados, se plantean la viabilidad técnica y la metodología de enriquecimiento de este conjunto de datos inicial en varios ejes complementarios:

1. La agregación de nuevos datos de la UNED, como tripletas RDF, que provengan de bases de datos insuficientemente integradas en las bases relacionales internas.

2. El enlace a *datastores* externos, permitiendo así consultas federadas con resultados más enriquecidos.

3. El enriquecimiento del *datastore* de la UNED con resultados de datos obtenidos de recursos documentales mediante técnicas de Minería de Textos. Donde dichos recursos documentales pueden obtenerse:

   a) Desde el mismo *datastore* propio, siguiendo las referencias a esos recursos que pueden encontrarse en los datos estructurales de la UNED (p.ej. guías de estudio, etc.).

[10] Almacén de Datos Enlazados RDF (Resource Description Framework).
[11] `http://unedata.uned.es`
[12] `http://portal.uned.es/portal/page?_pageid=93,39008560&_dad=portal`

b) Por otros medios de navegación guiada sobre los recursos UNED.

Este trabajo se enmarca dentro del Grupo de Investigación Docente de la UNED, Intelligent Systems for Learning[13] (ISLearning), y se centra en la explotación inteligente de los datos y recursos de la UNED: tecnológicamente, porque la escala del problema es relevante, e institucionalmente porque se pretende contribuir a la formalización de una base de conocimiento institucional que mejore alguna de las funcionalidades de esta universidad.

1.2. Objetivos del trabajo

En este trabajo nos centraremos en modelar y poblar un *datastore* mediante tripletas RDF con los datos disponibles sobre la oferta formativa de la UNED (Grados, Másteres y Curso de acceso), asignaturas en las que se divide esta oferta, su asignación al personal docente, y su división dentro de la estructura organizativa de la Universidad (departamentos y facultades o escuelas). Además, en una segunda fase se pretende un enriquecimiento del *datastore* generado de manera que, las asignaturas tengan enlazado un cierto número de palabras clave que puedan representar su contenido, con la intención de potenciar los posibles casos de uso de la colección de datos.

Con este propósito, se pretende presentar un modelo institucional de datos abiertos enlazados que sirva como propuesta de transferencia tecnológica, y que no sólo recoja el volcado de sus bases de datos relacionales, sino que permita, a diversos grupos de investigación, integrar incrementalmente resultados de analítica sobre recursos y procesos de la UNED.

Los objetivos marcados por este trabajo son los siguientes:

1. Construcción de un *datastore* de tripletas RDF con los datos sobre la estructura organizativa de la UNED, su personal académico y su oferta formativa y comprobar su respuesta funcional mediante consultas SPARQL. Dicho *datastore* se creará sobre un vocabulario propio, básico aunque suficiente para la gestión interna de estos datos.

   De manera general, el proceso de adaptación de los datos contenidos en la UNED a los principios fundamentales y las buenas prácticas de la Web Semántica y Datos Enlazados, constaría de una serie de fases:

   a) Análisis de los datos a modelar.
   b) Diseño y modelado de una ontología que de estructura a esos datos.
   c) Poblado automático de dicha ontología.
   d) Visualización y recuperación de la información: ya sea a través de la realización de consultas SPARQL a un *endpoint*[14], o mediante herramientas de visualización desarrolladas para la explotación y consulta.

2. Enriquecimiento documental del *datastore*.

   a) Detección de enlaces, desde el *datastore* construido en el punto anterior, a recursos UNED con información relevante aunque no tan estructurada. En particular, se pretende el diseño de un sistema de recopilación automática, en un único corpus, de todas las páginas sobre "Contenido" de las más de 3000 guías didácticas de asignaturas de la UNED publicadas en la Web.
   b) Estudio comparado de algoritmos de detección de términos y multitérminos singulares en fragmentos de este corpus (inicialmente asignaturas) respecto al corpus general o a subconjuntos del mismo, con implementación así mismo, de los procesos necesarios para su evaluación.

[13] `http://data.ia.uned.es/index.html`
[14] Un *endpoint* es un punto de conexión accesible a través de una red, desde el que se proporciona un servicio.

c) Poblado del *datastore* RDF inicial con estas colecciones seleccionadas de términos, ampliando el vocabulario básico para etiquetar las relaciones de los datos nucleares iniciales con estos otros complementarios.

3. Como objetivo secundario de investigación, se pretende realizar una pequeña revisión sobre la posibilidad de realizar mejoras en el vocabulario propuesto, de forma que se abra la posibilidad de efectuar consultas externas al *datastore*. Dichas mejoras en el vocabulario pueden ser objeto de discusión y uso conjunto con otras instituciones.

Los diferentes objetivos marcados en cada una de estas etapas serán desarrollados a lo largo del Capítulo 3.

1.3. Alcance y aplicaciones

Llevar a cabo el objetivo de una manera eficaz se podrá traducir en una serie de casos de uso que van, desde su publicación y enriquecimiento, hasta las posibles aplicaciones que se describen a continuación.

Una aplicación evidente de la consecución de los objetivos es la navegación, consulta y explotación de la información modelada (asignaturas, oferta formativa, profesorado, etc.). Por ejemplo, "qué profesores imparten qué asignaturas", "qué asignaturas contiene el segundo cuatrimestre de un determinado Grado", "en qué departamento trabaja cierto profesor", etc. Pudiendo escalar semánticamente en complejidad al nivel de inferencia que se desee y se permita.

Otras aplicaciones se derivan del enriquecimiento del *datastore* con palabras clave para cada asignatura; una de ellas se obtiene al complementar la información modelada con datos externos enlazados. Por ejemplo, con enlaces a DBpedia o a otros recursos educativos (sin ir más lejos, la UNED dispone de un amplio catálogo de cursos, materiales audiovisuales, trabajos de investigación, etc.).

Una primera aplicación puesta en marcha por el equipo de ISLearning consiste en el desarrollo de una interfaz de navegación a través de nubes terminológicas (nubes de *tags*). En una primera versión muestra la terminología relacionada con las asignaturas de un Grado previamente seleccionado, navegando por los distintos cuatrimestres que lo componen[15].

Además, disponiendo de los términos más representativos de cada asignatura, se puede recomendar a un alumno cursar unas titulaciones u otras en función de sus intereses. Del mismo modo, por ejemplo, también se puede orientar sobre qué asignaturas debería cursar un alumno que quiere matricularse de un Máster que cuenta una amplia oferta de éstas, basándose en sus intereses, objetivos o expediente académico. Otra aplicación posible de la nube de *tags* podría ser su utilización por parte de los coordinadores de Grado, de modo que revisen las asignaturas de un determinado Programa de Estudios y a partir de su terminología deteminen si hay solapamientos, carencias, etc., en contenidos.

Las posibilidades consecuencia de este trabajo crecerán considerablemente, a medida que la gran variedad de recursos de la Universidad vayan siendo conectados y publicados como Datos Enlazados, pudiendo tomar como referencia, si se desea, iniciativas de otras universidades que ya gozan de las ventajas fruto del acceso, reutilización y explotación de los Datos Enlazados.

[15] `http://data.ia.uned.es/nubeTagsDef/terminos_v7.html`

Capítulo 2

Estado de la cuestión

Este capítulo se ha dividido en dos secciones. En la primera de ellas, "Web Semántica, Datos Enlazados y Sistemas Educativos", se pretende tomar una panorámica del beneficio que aporta el uso de Datos Enlazados y de Tecnologías Semánticas dentro del marco universitario. Además, dado que nuestro *datastore* será enriquecido con palabras clave extraídas desde fuentes de texto libre, la segunda sección del capítulo, "Minería de textos", está dedicada a analizar este área de investigación y presentar las técnicas y funciones de ponderación que hemos considerado que mejor se adaptan a nuestro objetivo.

2.1. Web Semántica, Datos Enlazados y Sistemas Educativos

Esta sección se divide en varios apartados que pretenden discurrir progresivamente desde la revisión a los conceptos de Web Semántica y Datos Enlazados hasta su aplicación en el marco universitario en general, y en la UNED en particular, creando una visión global del estado de la cuestión en lo referente a este ámbito. Para ello se presentan a continuación los siguientes apartados:

1. Tecnologías Semánticas y Sistemas Educativos
   a) Web Semántica y Datos Enlazados, introducción a conceptos, principios y tecnologías.
   b) Sistemas Educativos en la Web e implicaciones de la introducción de las Tecnologías Semánticas.
2. Iniciativas, vocabularios y *datastores* en el entorno universitario
   a) Linked Universities como alianza de universidades que se beneficia de la utilización de Datos Enlazados y Tecnologías Semánticas.
   b) Otras iniciativas y proyectos de Datos Enlazados en el sector educativo, incluyendo otras ontologías y herramientas que no fueron revisadas en el apartado de Linked Universities.
   c) La Open University como paradigma en la utilización de proyectos basados en Tecnologías Semánticas y Datos Enlazados.
   d) La Universidad Nacional de Educación a Distancia, presentación e introducción a iniciativas semánticas.

2.1.1. Tecnologías Semánticas y Sistemas Educativos

Web Semántica y Datos Enlazados

Hoy en día existe una gran cantidad y diversidad de datos de diferentes temáticas en la Web, que proporcionan información disponible y fácilmente interpretable para todas las personas.

Sin embargo, esta información es inmensa y crece cada día, con lo que surge la necesidad de gestionarla de una manera eficaz y eficiente.

Por su parte, la WWW[1] presenta la información en un formato destinado a las personas como usuarios finales de ésta pero careciendo de semántica para los ordenadores. El objetivo de la Web Semántica es dotar de significado a los datos contenidos en las páginas Web, creando un entorno donde agentes *software* los recorren página a página, pudiendo, fácilmente, llevar a cabo sofisticadas tareas para los usuarios [6].

En [6] se presenta el concepto de la **Web Semántica** como una extensión de la Web actual en la cual se da un significado bien definido a la información, permitiendo mejorar la comunicación entre personas y computadores en la Web. Al igual que en la WWW las páginas se interconectan entre sí a través de hiperenlaces, en la Web Semántica los enlaces se establecen sobre los datos que contienen las distintas páginas y, del mismo modo, se presentan de la manera más descentralizada posible. La idea de un sistema complejo de información entrelazada en la Web aparece por primera vez en [4] de la mano de Tim Berners-Lee. Con este enfoque de la Web se obtienen estructuras más complejas de datos, estableciendo las reglas para realizar razonamientos, y utilizando la lógica para llevar a cabo inferencias sobre estos datos, proporcionando a los ordenadores un acceso más avanzado a la información. La Web Semántica se desarrolla bajo 6 principios fundamentales [33]:

1. Todo puede ser identificado mediante URIs[2]. Todos los recursos, incluyendo los vocabularios que los describen, son identificados mediante URIs.

2. Los recursos y enlaces pueden tener tipos que proporcionen mayor información a los agentes sobre los conceptos que manejan.

3. Debe existir tolerancia hacia la información parcial y hacia la perdida de ésta. Algunos de los recursos enlazados pueden dejar de existir con el tiempo, por lo que se debe persentar cierta tolerancia hacia esta carencia de datos.

4. No hay necesidad de una verdad absoluta. Serán los agentes los encargados de decidir el grado de verdad en función del contexto.

5. Tolera la evolución resolviendo ambigüedades y añadiendo nueva información, según sea necesario.

6. Diseño minimalista, sin estandarizar más de lo necesario.

Los recursos, sus propiedades y relaciones estarán dotados de una estructura que los defina y los integre. Tal y como se expone en [29], una **ontología** es una especificación explícita de una conceptualización. El término se toma prestado de la filosofía, donde se define como una descripción sistemática de la existencia. Para los sistemas basados en el conocimiento, lo que "existe" es exactamente aquello que puede ser representado.

Las ontologías se presentan como jerarquías de conceptos enmarcados en un dominio. Dichos conceptos, vienen definidos por "clases" y "atributos", presentando también "relaciones", y dando lugar en su conjunto a una red semántica de información. Cada uno de los individuos que forman parte de una clase se representa por medio de una "instancia" de ésta, y cada una de las relaciones entre individuos puede mostrar restricciones que se presentan en forma de "axiomas". En el ámbito de la Web Semántica, las ontologías son conjuntos de metadatos que proporcionan un vocabulario controlado de conceptos, compuestos de una semántica explícitamente definida y procesable por un equipo [39].

A partir de la noción de Web Semántica se consolida el concepto de **Datos Enlazados**, el cual se refiere a las buenas prácticas para la utilización de la Web con el fin de publicar y

[1] Worl Wide Web
[2] Universal Resource Identificator

conectar información a través de enlaces entre datos procedentes de diversas fuentes [10, 9]. Berners-Lee plantea en [5] una serie de principios para llevar a cabo una correcta generación de Datos Enlazados:

1. Utilizar las URIs para identificar las cosas.
2. Utilizar URIs HTTP.
3. Proporcionar información útil haciendo uso de estándares como RDF o SPARQL.
4. Incluir enlaces a otras URIs relacionadas, con el fin de ofrecer la oportunidad de descubrir nueva información.

La maduración de las tecnologías de la Web Semántica, junto al crecimiento de la información publicada bajo los principios de Datos Enlazados, ha desembocado en la aparición de una nueva concepción conocida como Web de Datos [10]. El W3C[3] trabaja en la creación de recomendaciones[4] que den soporte a dicha Web de Datos, con el objetivo de permitir a los computadores realizar un trabajo más útil y mejorando las interacciones y el intercambio de información en la Web. Además, las Tecnologías Semánticas permiten la creación de almacenes de datos, vocabularios y reglas para gestionar la información.

Dos tecnologías muy empleadas para dotar de estructura a los datos son XML[5] y RDF. En XML los usuarios pueden dar una estructura arbitraria a sus documentos pero sin dotarlos de una semántica sobre la que llevar a cabo posibles inferencias. Para realizar esta tarea se hace uso de RDF, que codifica la información en tripletas cuyos componentes son sujeto, verbo y predicado. Cada tripleta presenta información sobre algo (sujeto) que tiene una propiedad (predicado) con cierto valor (objeto). Sujeto, predicado y objeto son referenciados mediante una URI. Las URIs forman documentos de conceptos relacionados y dotados de un significado, siendo fácilmente accesibles por cualquier usuario [6].

Por otro lado, dos documentos distintos pueden hacer referencia a un mismo concepto utilizando identificadores diferentes. Un programa (agente) que recorra ambos documentos y obtenga o compare su información, debe conocer que ambos identificadores se refieren al mismo concepto, presentando consecuentemente el mismo significado. Para llevar a cabo este propósito se hace uso de las ontologías [6], pudiéndose realizar sobre éstas tareas de reutilización, modificación e integración de uno o varios vocabularios.

El lenguaje RDF[6] es la recomendación del W3C para representar recursos y sus relaciones en la Web. Utiliza URIs y literales (datos) para la construcción de grafos dirigidos y etiquetados mediante la combinación de tripletas RDF. Existen diferentes vocabularios RDF en función de la naturaleza de los recursos que se quieren describir. Para una correcta comunicación entre distintas máquinas, éstas deben conocer el mismo vocabulario.

Con el fin de estructurar correctamente la información, se utilizan modelos ontológicos de datos que permiten definir tanto la estructura, como las restricciones existentes entre éstos. Aquí cabe una especial mención a vocabularios como RDFS[7], OWL[8] y OWL2[9] que permiten crear clases, instancias y propiedades.

Como medio para recuperar datos que se encuentran en formato RDF se utiliza de manera general el estándar SPARQL[10] [53]. SPARQL puede realizar consultas a través de diversas fuentes

[3] World Wide Web Consortium: es una comunidad internacional liderada por Tim Berners-Lee, cuyos miembros trabajan para desarrollar los estándares que definen la Web. `http://www.w3c.es`

[4] Recomendaciones W3C: `https://www.w3.org/TR/tr-technology-stds`

[5] eXtensible Markup Language

[6] `https://www.w3.org/RDF`

[7] `https://www.w3.org/TR/rdf-schema`

[8] `https://www.w3.org/TR/2003/PR-owl-features-20031215`

[9] `https://www.w3.org/TR/owl2-primer`

[10] `https://www.w3.org/TR/rdf-sparql-query`

de datos y tiene la capacidad de consultar grafos que cumplan con ciertos patrones tanto obligatorios como opcionales, expresados en forma de conjunciones y disyunciones. SPARQL puede devolver resultados a una consulta tanto en forma de objetos, predicados o sujetos (recursos o literales), como en forma de grafos RDF.

Sistemas Educativos

Los Sistemas Educativos han ido adquiriendo cada vez más popularidad con el paso del tiempo, gracias en gran parte a la evolución y expansión de la WWW. Los beneficios de estos sistemas son claros: la independencia de la clase y la independencia de la plataforma. La utilización del mismo material didáctico por miles de estudiantes de todo el mundo junto a la explotación de aplicaciones adaptativas inteligentes[11], se han convertido en requisitos fundamentales en este tipo de sistemas, ya que los estudiantes a distancia por lo general trabajan por su cuenta y no es fácil obtener una asistencia inteligente además de personalizada [14].

Pero en los últimos tiempos dichos sistemas han ido un paso más allá, incorporando los beneficios proporcionados por las Tecnologías Semánticas. Los Sistemas Educativos basados en la Web semántica son la nueva generación de Sistemas Educativos que evolucionan en Sistemas Educativos más personalizables, flexibles e inteligentes. El objetivo principal de estos sistemas es utilizar los recursos disponibles en la Web a través de tecnologías basadas en estándares para lograr que cualquiera pueda aprender en cualquier lugar, en cualquier momento [8].

Las ventajas derivadas de la utilización de Datos Enlazados en el sector educativo son, según [23]:

- Mayor facilidad en la navegación y acceso a los recursos educativos debido a la utilización de metadatos. Algunas aplicaciones utilizan los Datos Enlazados para implementar interfaces de navegación y visualización de recursos educativos.
- La interconexión automática entre recursos procedentes de distintas fuentes, lo cual facilita la recomendación y el descubrimiento de nuevos recursos que pueden resultar de interés para el usuario. Por ejemplo, sugerir recursos que han sido creados por los mismos autores que los recursos que se están recuperando.
- Personalización y aprendizaje social: un sistema orientado a la personalización puede recomendar al alumno recursos de aprendizaje y actividades, basándose en los intereses de éste, su historial, objetivos y logros. Además, el alumno puede compartir su experiencia de aprendizaje e interactuar con otros alumnos mediante conexiones sociales.

2.1.2. Iniciativas, vocabularios y *datastores* en el entorno universitario

Linked Universities

Linked Universities[12] es una alianza de universidades europeas dedicadas a exponer sus datos públicos como Datos Enlazados. En su portal web se puntualiza que, actualmente, existen pocas universidades que lleven a cabo estas buenas prácticas. Además, se suma el hecho de que muchas de estas iniciativas suelen estar desconectadas por lo que, se desaprovecha parte del potencial y de los beneficios que la Web Semántica puede desarrollar en el ámbito interuniversitario. Linked Universities pretende ser un espacio de colaboración donde las instituciones y los individuos involucrados en la creación y publicación de datos vinculados a universidades puedan describir, compartir y reutilizar vocabularios y prácticas comunes.

Las universidades que forman dicha alianza son la Universidad de Bristol en Inglaterra, la Universidad de Southampton en Inglaterra, la Universidad de Ege en Turquía, la Universidad

[11] Sistemas de Tutorías Inteligentes y Sistemas Hipermedia Adaptativos
[12] `http://linkeduniversities.org`

de Aalto en Finlandia, el Consejo de Investigación Nacional Italiano, la Universidad de Münster en alemania, la Universidad de Charles, la Universidad de Aristóteles en Grecia, la Universidad Pompeu Fabra en Barcelona, y la Open University en el Reino Unido.

Linked Universities promueve el uso de vocabularios para la descripción de Datos Enlazados relacionados con el ámbito universitario: en su sitio web se presentan distintos vocabularios categorizados según la finalidad hacia la que se orientan:

- Vocabularios orientados a cursos:
  - *MLO - Metadata for Learning Opportunities*: proporciona un estándar para la publicación de cursos y su posterior recuperación por futuros estudiantes [13].
  - *XCRI-CAP - XCRI Course Advertising Profile*: el objetivo de este vocabulario es disponer de un catalogo de cursos, con el fin de permitir un sencillo intercambio de información entre varias universidades de Reino Unido[14].
  - *TEACH - Teaching Core Vocabulary*[15]: es un vocabulario que permite establecer relaciones con un curso, como por ejemplo: qué estudiantes o profesores tiene, qué material es necesario, dónde se imparte, etc.
- Vocabularios orientados a la descripción de la universidad como una organización:
  - *AIISO - Academic Institution Internal Structure Ontology*: es un vocabulario que proporciona clases y propiedades para describir la organización interna de una institución académica[16].
  - La ontología *Bowlogna*[17]: la ontología *Bowlogna* está definida a partir de los términos relacionados con el Proceso de Bolonia, que tiene como objetivo introducir un marco común de títulos transparentes y comparables, el cual garantice el reconocimiento de competencias, cualificaciones y conocimientos de los ciudadanos de la Unión Europea. Dicho proceso tiene los siguientes objetivos:
    - Adaptar las universidades al Espacio Europeo de Educación Superior[18].
    - Adaptar los títulos para fomentar su reconocimiento.
    - Adaptar y regular la estructura de las titulaciones de enseñanza superior.
    - Establecer un sistema común de créditos.
    - Promover la movilidad (Erasmus).
    - Promover la cooperación europea en la garantía de la calidad.
    - Promover el aprendizaje permanente, fomentando la educación en adultos.
  - Core Organization Ontology[19]: fue originalmente creada con el objetivo de publicar datos sobre la estructura organizativa del gobierno de Reino Unido, pero pronto se amplió a otros dominios como el universitario. La ontología no proporciona una información completa de la organización, pero sí suministra unos conceptos básicos que después se podrán ir ampliando según sea necesario.
- Vocabularios orientados a la descripción de publicaciones académicas y comunidades de investigación:

[13] `ftp://ftp.cenorm.be/PUBLIC/CWAs/e-Europe/WS-LT/CWA15903-00-2008-Dec.pdf`
[14] `http://eprints.worc.ac.uk/649/1/EUNISuwxcri.pdf`
[15] `http://linkedscience.org/teach/ns`
[16] `http://vocab.org/aiiso`
[17] `http://diuf.unifr.ch/main/xi/bowlogna`
[18] `http://www.eees.es`
[19] `http://epimorphics.com/public/vocabulary/org.html`

- *BIBO - The Bibliographic Ontology*[20]: proporciona conceptos y propiedades para describir citas y referencias bibliográficas en la Web Semántica.
- *VIVO Ontology*[21]: además de un vocabulario, proporciona un software de código abierto que potencia el trabajo colaborativo entre distintas instituciones. Este proyecto está orientado a la Web Semántica y se completa con información sobre el interés, actividad y logros de los investigadores de una institución. Esto permite la edición, búsqueda, navegación y visualización de su actividad académica, mostrando sus áreas de experiencia, credenciales académicas, redes de trabajo y toda la información sobre sus publicaciones y proyectos.

- Otros vocabularios:
  - *Mathematics Subject Classification*[22]: a través de la utilización del lenguaje SKOS[23] trata de implementar una clasificación de las distintas áreas que integran las matemáticas.
  - *W3C Ontology for Media Resources*[24]: introduce un vocabulario que pretende unificar las diferentes descripciones que presentan los recursos multimedia, proporcionando un conjunto básico y común de propiedades descriptivas.
  - *Linked Science Core Vocabulary (LSC)*[25]: LSC es un vocabulario ligero que proporciona términos que permiten a editores e investigadores describir recursos científicos incluyendo elementos de investigación, su contexto e interconexión.
  - *The Common European Research Information Format Ontology Specification (CERIF Ontology)*[26]: proporciona conceptos básicos y propiedades con el fin de describir, mediante datos semánticos, información sobre investigación.

Otras iniciativas y proyectos de Datos Enlazados en el sector educativo

Además de Linked Universities, existen varios proyectos e iniciativas relacionadas con la creación y publicación de Datos Enlazados Abiertos por parte de las instituciones educativas. Aquí haremos un breve repaso de las que hemos considerado más relevantes.

- Linked Education[27] es una plataforma web abierta destinada a facilitar y promover el intercambio de datos y recursos educativos mediante la utilización de Datos Enlazados. Además, proporciona un espacio donde investigadores y profesionales en los campos vinculados a las Tecnologías Semánticas y a la educación a través de la Web pueden estudiar y compartir conjuntos de datos, esquemas y aplicaciones, identificando cuáles son las mejores prácticas así como los posibles enlaces entre los distintos recursos.

- LinkedUp[28] es un proyecto dirigido por la Universidad de Leibniz en Hannover que tiene como objetivos: la creación de sistemas que integran Datos Enlazados Abiertos a gran escala por parte de instituciones y organizaciones educativas, la recopilación de Datos Enlazados relevantes de contenido educativo para su explotación por parte de terceros, la evaluación de aplicaciones de datos abiertos teniendo en cuenta aspectos educativos y la promoción de las tecnologías de datos abiertos dentro del ámbito educativo.

[20] `http://bibliontology.com`
[21] `http://vivoweb.org`
[22] `http://msc2010.org/resources/MSC/2010/info`
[23] SKOS es un vocabulario que proporciona un modelo para representar la estructura básica y el contenido de esquemas conceptuales: `https://www.w3.org/TR/skos-reference`
[24] `https://www.w3.org/TR/mediaont-10`
[25] `http://linkedscience.org/lsc/ns`
[26] `http://eurocris.org/ontology`
[27] `http://linkededucation.org`
[28] `https://linkedup-project.eu`

- LAK Dataset[29] publica versiones comprensibles a nivel máquina de fuentes de investigación pertenecientes a las comunidades de Análisis de Aprendizaje y Minería de Datos Educativos. Esta plataforma tiene el objetivo de facilitar la investigación, el análisis y el desarrollo de aplicaciones inteligentes en este área.

Sin salirnos del marco educativo existen herramientas que facilitan la explotación y publicación de Datos Enlazados, como por ejemplo:

- GNOSS[30] permite la interconexión, mediante Tecnologías Semánticas, entre personas, empresas y grupos u organizacines en función de sus intereses.

  Una de las comunidades dentro de GNOSS que resulta de especial interés es GNOSS Universities[31], que es una plataforma social y semántica para la creación de sistemas de conocimiento sociales que se basa en la publicación y explotación de Datos Enlazados Abiertos en el ámbito de la universidad, mejorando la gestión de ésta y los procesos de aprendizaje organizacional.

- CKAN[32] es una herramienta de código abierto para crear sitios web donde administrar y publicar colecciones de datos abiertos. Es comúnmente utilizada por gobiernos, instituciones de investigación y otras organizaciones que trabajan con grandes cantidades de datos.

A continuación, se presentan otras ontologías también utilizadas en el área educativa:

- HERO es una ontología diseñada con la intención de establecerse como referencia en el marco de la educación superior. En [68] se presenta el proceso de creación de la ontología de este tipo desde su especificación hasta su evaluación.

- SWRC es una ontología diseñada con el fin de modelar comunidades de investigación y aquellos conceptos relevantes relacionados con éstas [62].

La Open University

La Open University[33] es la mayor institución académica del Reino Unido y la universidad pionera a nivel mundial en aprendizaje a distancia. A día de hoy cuenta con más de 170000 estudiantes repartidos por todo el mundo.

La Universidad cuenta con varios sitios web con contenido de libre acceso, entre los que se encuentra un canal de vídeos en YouTube[34], otro de *podcasts* en AudioBoom[35], y la plataforma OpenLearn[36] de cursos en abierto.

Pero además del contenido multimedia la Open University ha creado una gran cantidad de recursos educativos a lo largo de los últimos 40 años. La utilización de todos estos recursos junto con la generación de otros nuevos son factores clave en el desarrollo efectivo de cursos nuevos. Así, en la Open University la distribución y reutilización de conocimiento es un reto donde los Datos Enlazados juegan un papel importante [20]. Esta institución forma parte de la alianza de universidades Linked Universities.

[29] http://lak.linkededucation.org
[30] urlhttp://www.gnoss.com
[31] http://www.gnoss.com/comunidad/gnossproducts/recurso/gnoss-universities-construyendo-ecosistemas-de-con/5fb9e264-c7be-486d-a492-fa798e881393
[32] https://ckan.org/
[33] http://www.open.ac.uk
[34] https://www.youtube.com/user/TheOpenUniversity
[35] https://audioboom.com/search?utf8=%E2%9C%93&q=the+open+university
[36] http://www.open.edu/openlearn

Gran cantidad de la información almacenada en diversos sitios web de la organización se reune en una única ubicación[37] común y de acceso abierto, gracias al proyecto LUCERO[38]. En el sitio web `http://data.open.ac.uk` se proporciona un acceso a los Datos Enlazados de la Universidad. En esta plataforma los datos vienen clasificados en seis conjuntos bien diferenciados:

1. Recursos educativos abiertos, como por ejemplo los cursos de OpenLearn.
2. Producción científica de la Universidad, con metadatos para llevar a cabo investigaciones abiertas en línea.
3. Medios de comunicación social, con contenido alojado en redes sociales como YouTube o AudioBoom.
4. Datos organizativos, recolectados de los repositorios internos de la Universidad y que se hacen públicos como Datos Enlazados Abiertos.
5. Datos enlazados de proyectos de investigación.
6. Metadatos y documentación, espacio de datos dedicados a la descripción y documentación de los distintos esquemas de Datos Enlazados de la Universidad.

Todos estos Datos Enlazados son accesibles, reutilizables y explotables, facilitando a la Open University aprovechar toda la información que estos proporcionan mediante la utilización de diversas aplicaciones desarrolladas con ese propósito. En [20, 24, 67] se presentan algunos ejemplos de estas aplicaciones entre los que se encuentran los siguientes:

- El sitio web de cursos OpenLearn, a través de consultas SPARQL, recupera una lista de cursos y títulos con toda la información relacionada.
- La Unidad de Servicios Estudiantiles de la Open University hace uso de los Datos Enlazados para actualizar la lista de cursos disponibles.
- Una aplicación en el canal de YouTube recupera información de los Datos Enlazados para obtener tanto cursos y títulos, como otros contenidos educativos abiertos relacionados con el contenido visualizado.
- El motor DiscOU [21] que mediante consultas a la DBpedia[39] hace recomendaciones de material educativo de contenido similar al de algunos recursos *online* externos a la Universidad como por ejemplo, una página web, o un vídeo visitado en la BBC.
- Una aplicación de móvil orientada a mostrar la oferta de títulos y cursos a los estudiantes de la Universidad. El estudiante introduce un tema que resulte de su interés y la aplicación le muestra todos los cursos relacionados con ese tema junto a otros recursos como vídeos, *podcasts*, etc.
- Buddy Study combina la red social Facebook con el servicio de Datos Enlazados de la Universidad, con el objetivo de sugerir compañeros de aprendizaje según el perfil del estudiante y los posibles cursos que podrían llevar a cabo juntos.
- El sistema de búsqueda experta de la Open University permite identificar gente experta en un dominio de interés, objetivo importante para toda empresa pero también para las universidades, donde se realizan colaboraciones interdisciplinares entre investigadores.
- Otras aplicaciones orientadas a tareas de investigación, como por ejemplo, la evaluación de su calidad en las distintas universidades.

[37] `http://data.open.ac.uk`
[38] Linking University Content for Education and Research Online
[39] `http://es.dbpedia.org/`

La Universidad Nacional de Educación a Distancia

En agosto de 1972 nace oficialmente la UNED. En un principio cuenta únicamente con dos vicerrectorados, uno de Humanidades y otro de Ciencias. Las unidades didácticas eran gratuitas y se enviaban a los estudiantes por correo postal. Desde entonces, la Universidad ha crecido enormemente llegando a estudiantes de todas las provincias de España e incluso del extranjero. Actualmente, cuenta con más de 260000 alumnos, más de 1000 docentes, 11 facultades y escuelas, 2 institutos de investigación, 1 escuela internacional de doctorado, cerca de 100 estudios oficiales entre Másteres y Grados, más de 100 grupos de investigación que cuentan con gran cantidad de publicaciones, cursos de formación permanente, un Centro de Medios Audiovisuales, además de una vasta colección recursos y materiales educativos.

Su oferta educativa se estructura en:

- Estudios oficiales:
  - Grados
  - Grados Combinados
  - Másteres EEES
  - Doctorados EEES y Planes Antiguos
  - Licenciaturas/Diplomaturas/Ingenierías
- Estudios de Acceso a la Universidad
- Formación Permanente
  - Formación Permanente y Profesional
  - Extensión Universitaria
  - Cursos de Verano
  - CUID (Centro de Idiomas Digital y a Distancia)
  - UNED Sénior
  - UNED Abierta

La UNED, a diferencia de la mayoría de las universidades españolas, ofrece la modalidad de estudio a distancia. Para ello hace uso de las nuevas tecnologías mediante cursos virtuales en Internet a través de la plataforma ALF, apoyándose además en las tutorías presenciales en sus centros asociados. Así mismo, para completar su oferta didáctica ofrece multitud de contenidos variados y completos a través de la televisión educativa y los programas de radio.

Cada curso dispone de una página web en la que se presenta una orientación general sobre éste: presentación, competencias, perfil de ingreso, guía de estudios, normativa, salidas profesionales, etc. Dentro del apartado guía de estudios se puede encontrar una lista con las asignaturas que se cursarán y dentro de cada asignatura se puede acceder a los subapartados: presentación, contextualización, requisitos previos, resultados del aprendizaje, contenidos de la asignatura (de especial relevancia en este trabajo), Equipo Docente, metodología, etc.

La UNED dispone, además, de un Centro de Medios Audiovisuales[40] (CEMAV) que ofrece gran variedad de soportes y formatos con el fin de ser un recurso útil tanto a alumnos como a profesores e investigadores.

El CEMAV diseña y produce contenidos audiovisuales para distintos canales de difusión: Radio, TV, Internet y UNED Editorial.

Por otro lado, la UNED utiliza varias plataformas para la difusión a través de Internet tanto del contenido producido por el CEMAV, como de los actos realizados en la sede Central (Congresos, Videoconferencias, etc.):

[40]CEMAV: `http://portal.uned.es/portal/page?_pageid=93,773691&_dad=portal&_schema=PORTAL`

- TeleUNED[41]: como plataforma multimedia a través de Internet para la difusión de las actividades académicas, docentes, culturales e informativas de la Universidad. En la actualidad, TeleUNED cuenta con más de 25.000 archivos de producción audiovisual y desde esta plataforma se puede consultar la programación televisiva y radiofónica diaria, semanal o mensual, además de acceder en línea a los contenidos.
- CanalUNED[42]: mediateca con miles de horas en series, audios y vídeos disponibles para su visualización y descarga.
- RTVE-UNED[43]: repositorio desde donde se puede visualizar o descargar material audiovisual de la UNED difundido a través de RTVE (TVE Internacional, La 2, Radio 3, REE y Radio 5) .
- Youtube-UNED[44]: canal de youtube para la reproducción de emisiones audiovisuales que cuenta con material audiovisual que se clasifica en los siguientes canales: UNED Conferencias, UNED Documentos, UNED Entrevistas, UNED Cursos, UNED Cursos MOOC/COMA, UNED Radio y UNED INTECCA.

Además de los canales que se han visto hasta ahora, la UNED ofrece acceso a más material a través de las siguientes plataformas: UNED Abierta[45], AVIP[46], OpenCourseWare[47], e-spacio[48] y resto de titulaciones y actividades docentes que forman parte del conjunto de recursos cerrados.

Muchos de los contenidos audiovisuales citados en este subapartado podrían ser de gran ayuda, actuando como materiales complementarios en el estudio de las distintas asignaturas ofertadas por los distintos cursos que imparte la Universidad.

Así mismo, hemos visto que la UNED dispone de una gran cantidad de recursos que pueden ser objeto de enlaces que permitan sacar la máxima rentabilidad a la información albergada por la Universidad. Con este objetivo ha nacido el proyecto UNEDATA (`http://unedata.uned.es`).

UNEDATA es una herramienta de descubrimiento de investigación centrada en permitir la colaboración entre científicos de todas las disciplinas. El proyecto se enmarca en el marco del Laboratorio de Innovación en Humanidades Digitales[49] (LiNHD), el cual pretende aumentar el impacto de las humanidades en el panorama científico global.

Dicho proyecto permitirá enlazar los datos existentes en la UNED de forma que su aprovechamiento sea más eficaz, incluso pudiendo llegar a enlazar información con otras instituciones y universidades. Haciendo uso, por supuesto, de las tecnologías y principios de la Web Semántica y los Datos Enlazados.

UNEDATA en su primera fase está centrado en enlazar datos relacionados con la investigación, cuya procendencia emana desde fuentes estructuradas, y en este caso en particular, de la base de datos de investigación científica, la cual contiene información sobre proyectos, grupos de investigación, cátedras, institutos, profesores, etc. Sin embargo, UNEDATA tiene como finalidad la creación de una universidad abierta, que se enmarca dentro de un amplio proyecto que actualmente se encuentra comenzando su camino.

Aunque a día de hoy y en su primera fase, el proyecto UNEDATA presenta algunas carencias en cuanto a población de datos y establecimiento de un *endpoint* accesible y consultable, con este proyecto la UNED se pretende sumar a la solución que brinda la utilización de Datos Enlazados dentro del contexto universitario, al igual que ya lo hacen otras universidades a nivel

[41] TeleUned: `https://canal.uned.es/`
[42] CanalUned: `https://canal.uned.es/`
[43] RTVE-UNED: `http://rtve.es/uned`
[44] Youtube-UNED: `https://www.youtube.com/user/uned?gl=ES&hl=es`
[45] UNED Abierta: `https://unedabierta.uned.es`
[46] AVIP: `https://www.intecca.uned.es/portalavip/plataformaAVIP.php`
[47] OpenCourseWare: `http://ocw.innova.uned.es/ocwuniversia/`
[48] e-spacio: `http://e-spacio.uned.es/`
[49] `http://linhd.es`

estatal como la Universidad Pompeu Fabra[50] (miembro de la alianza Linked Universities), la Universidad Pablo de Olavide[51], o la Universidad de Deusto[52], entre otras.

2.2. Minería de Textos

El problema de la búsqueda de información en textos se remonta a la década de los 50, donde Hans Peter Luhn [38] plantea una serie de pasos con el fin de configurar un Sistema de Búsqueda Literaria.

Más tarde, Salton y McGill [58] presentan un método de representación de los documentos en el espacio vectorial (figura 2.1), donde cada documento viene representado por un vector, con tantas componentes como términos diferentes tenga la colección. Una función se encarga de asignar pesos a los términos, de manera que las componentes de cada vector definan a éste en el espacio. Este tipo de funciones se conocen como funciones de ponderación o *term weighting functions*, en su expresión en inglés.

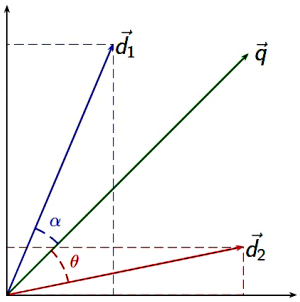


Figura 2.1: Modelo de espacio vectorial

Haciendo uso de medidas de similitud entre vectores, se establecen entre otras, las siguientes aplicaciones de interés en Recuperación de Información[53] (IR):

- Recuperación de documentos a partir de consultas: en ocasiones se dispone de textos muy largos que se quieren recuperar de manera efectiva a través de consultas con muy pocos términos. En algunos casos se recurre a la expansión automática de consultas para tratar de ampliar la cobertura de la recuperación [44, 56, 54, 66].

- Clasificación automática de textos o categorización: asume la existencia de distintas clases de documentos y las asigna a cada texto nuevo que se pretenda categorizar, de manera supervisada y tras una fase de entrenamiento en la que se encuentra un modelo para cada clase [41, 42, 7].

- *Clustering* de documentos: esta técnica es utilizada para agrupar documentos donde se trata de maximizar la similitud intraclase, maximizando a su vez la similitud intraclase. Su objetivo puede ser el de mejorar la eficiencia y efectividad en la recuperación de información, o el de determinar la estructura que gobierna la literatura dentro de un campo. Es una técnica de clasificación no supervisada que consiste en crear grupos en base al cálculo del grado de asociación entre elementos. A diferencia de la categorización, en el *clustering* no existen unas clases predefinidas, sino que éstas se definen una vez se genera la agrupación de manera automática [60, 1, 48].

[50] `https://data.upf.edu/es/main`
[51] `https://datos.upo.gob.es`
[52] `https://datahub.io/dataset/deustotech`
[53] La Recuperación de Información consiste en encontrar material (generalmente documentos) de naturaleza no estructurada (normalmente texto) dentro de grandes colecciones, con el fin de satisfacer una necesidad de información [40].

En este trabajo no se aplicarán similitudes entre vectores de representación de documentos, sino que nos centraremos en la ordenación y ponderación de los términos presentes en éstos según lo importantes que sean para el mismo, y así poder después extraer aquellos que resulten más significativos para el contenido del documento. Dicho lo cual, en [57] se hace referencia a dos cuestiones fundamentales a tener en cuenta:

1. ¿Cómo deben ser las unidades de texto mínimo dentro de los documentos? En muchas ocasiones un concepto representativo se forma por la unión de dos o más términos (tokens), los cuales, de ir por separado, perderían la semántica de la unión y, por consiguiente, dejarían de ser una representación adecuada del contenido del documento.

2. ¿Una adecuada asignación de pesos a términos puede distinguir aquellos que resultan más representativos de los menos importantes dentro del contenido de un documento? Los pesos no sólo influyen en la determinación de la significancia de los términos en el documento, sino en la representación del documento en el espacio vectorial. Según [34] ha habido mucha investigación sobre técnicas de pesado terminológico, pero poco consenso sobre cuál es la mejor porque, fundamentalmente, depende en gran medida del objetivo del pesado. No es lo mismo encontrar un buen pesado para un proceso de Recuperación de Información, donde lo que se quiere es buscar la importancia de un término respecto del contenido de un documento, que buscar un buen pesado para un proceso de clasificación o de *clustering* de documentos, donde lo que se busca precisamente no es la diferencia entre documentos, sino la similitud entre algunos de ellos. Además, en dicho texto, también se indica que un esquema de pesado debe de estar formado por tres componentes: el factor local, el factor global y el factor de normalización[54].

Existen varias técnicas que han demostrado ser efectivas en asignación de pesos en el área de IR. En [57, 50] se puede encontrar un análisis y comparación de distintos enfoques. En muchos casos se emplean estas mismas funciones para problemas de clasificación o *clustering* de documentos, como para con el estándar de facto TF-IDF, pero ya se ha indicado que esto puede no ser del todo correcto, al no buscarse el mismo objetivo en dichas aplicaciones. A lo largo de esta memoria se volverá a discutir este tema más detalle.

Funciones de ponderación

En [50] se enuncia que existen tres factores a tener en cuenta en el momento de evaluar un término: el factor local, el factor global y la normalización. Además, en [57, 50] se pueden consultar varias funciones englobadas en cada uno de estos factores. En este proyecto, nos centraremos en las funciones que, en la literatura, han demostrado ser más efectivas y aquellas que mejor pueden adaptarse a nuestro corpus.

El **factor local** es la medida de representación que se asigna a un término teniendo como referencia, únicamente, el documento al que pertenece. Existen varias funciones que determinan el factor local (ver [57, 50]) y que pueden resultar más o menos interesantes dependiendo del objetivo a alcanzar, pero la función local clásica es **TF** y viene definida en la ecuación 2.1. En TF se considera que los términos serán más o menos importantes en función de su frecuencia de aparición en el documento. Luhn en [38] asumía la idea de que cuanto mayor fuera el número de apariciones de un término en un documento, más importante sería dicho término en éste.

$$TF(t,d) = f_{t,d} \tag{2.1}$$

Donde $f_{t,d}$ es la frecuencia para el término t en el documento d.

[54]El factor de normalización se utiliza para corregir las discrepancias existentes entre las longitudes de los documentos, de manera que estos puedan ser comparados [50]. Este factor no es de especial relevancia para este trabajo, debido a que no se llevarán a cabo comparaciones entre documentos en el espacio vectorial

Esta función TF suele ser empleada en combinación con otro factor global, ya que, por sí sóla, favorece a aquellos documentos más largos y también a las palabras que presentan grandes frecuencias [34]. Sin embargo, en ciertas ocasiones, puede ser preferible utilizarla sin combinarla con otros factores. Por ejemplo, en textos muy cortos y en la asignación de pesos a consultas, donde los términos suelen aparecer una o dos veces [50].

El **factor global** mide la importancia de un término de un documento en función de su presencia en el resto de la colección. Se asume que a mayor cantidad de documentos donde dicho término esté presente, menos representativo será éste de su documento. La función clásica utilizada para medir el factor global de un término es **IDF**, también conocida como "frecuencia inversa de documento", y viene definida en la ecuación 2.2. En la literatura, se suele utilizar esta función en combinación con TF (dando lugar a TF-IDF), aunque también puede ir sóla [55].

$$IDF(t) = log\left(\frac{N}{df(t)}\right) \tag{2.2}$$

Donde N viene dado por la cantidad de documentos del corpus, y $df(t)$ por la frecuencia de documento, es decir, el número de documentos en los que aparece el término t.

Según [50] existen dos importantes razones por las que utilizar el **factor de normalización** en la asignación de pesos en términos:

1. Los documentos largos tienden a repetir los términos, por lo que estos términos tendrán una frecuencia mayor.

2. Los documentos largos también tienen mayor cantidad de términos diferentes, por lo que al realizar una consulta terminológica sobre la colección, es más sencillo que dicha consulta comparta términos con un documento largo, antes que con uno corto.

Hasta ahora, se han descrito técnicas sencillas de asignación de pesos, basadas en el factor local y el factor global de manera independiente. Sin embargo, lo habitual es encontrar combinaciones de funciones que unifiquen ambos factores y con las que se han venido obteniendo mejores resultados durante los últimos años. En este trabajo nos vamos a centrar también en dos técnicas con gran relevancia: TF-IDF y la Divergencia de Kullback-Leibler (KLD).

- **La función TF-IDF:** es una de las técnicas más utilizadas para asignación de pesos en IR [59, 2]. Dicha función es definida en la ecuación 2.3.

$$TF\text{-}IDF(t,d) = f_{t,d} \cdot log\left(\frac{N}{df(t)}\right) \tag{2.3}$$

  Donde $f_{t,d}$ es la frecuencia para el término t en el documento d, N viene dado por la cantidad de documentos del corpus, y $df(t)$ por el número de documentos de éste en los que aparece el término t.

  TF-IDF combina el factor local TF y el factor global IDF como componentes para la asignación de un peso a cada término. Por lo tanto, otorga importancia a un término en función de la frecuencia con la que éste se presenta en el documento, añadido a lo específico que sea en la colección. Es decir, un término será más importante para un documento en proporción directa al número de apariciones tenga en éste, y en proporción inversa a la cantidad de documentos que contengan a dicho término.

  El esquema TF-IDF, se ha considerado a menudo un método empírico con muchas variaciones posibles [2]. Algunos autores han realizado interpretaciones probabilísticas de esta técnica de recuperación de información, de manera que proporcionan una explicación estadística a esta función que históricamente ha sido considerada heurística [30, 2].

- **La función KLD:** en [15] se enuncia una función de pesado cimentada a partir de la Divergencia de Kullback-Leibler [36] (KLD). La función **KLD** se basa en las diferencias existentes entre las distribuciones de los términos del documento a valorar y las distribuciones de los términos contenidos en el corpus. La ecuación 2.4 asigna un peso a cada término t contenido en el documento d en función de su aportación a la entropía relativa existente entre ambas distribuciones (entre el documento y el corpus).

$$KLD(t,d) = P(t,d) \cdot log\left(\frac{P(t,d)}{P(t,C)}\right) \tag{2.4}$$

  Donde $P(t,d)$ viene dado por el valor la probabilidad obtenida tras la división de la frecuencia de aparición del término t en el documento d, por la suma de frecuencias de cada uno de los términos del documento d, y $P(t,C)$ viene determinado por la probabilidad obtenida tras la división de la frecuencia de aparición del término t en el corpus de todos los documentos C, por la suma de frecuencias de cada uno de los términos de C.

  Con KLD se asignarán los pesos a los términos de manera que, aquellos que aparezcan con bastante probabilidad en un documento y con poca probabilidad en el corpus serán los que más valor obtengan. Esta función confiere un valor a cada término en dependencia con la divergencia entre la probabilidad con la que éste aparece en el documento y la probabilidad con la que aparece en la colección (por ello, KLD se confiere como la mejor candidata en la extracción terminológica que se desea llevar a cabo en este trabajo). Así mismo, de manera análoga a la combinación de factor local y factor global que se llevaba a cabo con TF-IDF, KLD también tiene en cuenta la distribución local y global del término para realizar la ponderación de éste.

  Desde su introducción en [15], KLD se ha venido utilizando tanto para asignar y recalcular pesos (p. ej. [3, 16, 32]), como para llevar a cabo otras funciones en IR, tales como encontrar términos para realizar expansión de consulta, o calcular distancias en tareas de clasificación y *clustering* de documentos (p. ej. [7, 48]).

Capítulo 3

Metodología, diseño y desarrollo de la experiencia

3.1. Construcción de un *datastore* de tripletas RDF

3.1.1. Dominio de los datos a modelar

Uno de nuestros objetivos es la representación de información estructurada perteneciente a la Universidad mediante Datos Enlazados. Para ello, emplearemos un vocabulario propio, básico y suficiente, que permita el acceso automatizado a los recursos analizables. Nos centraremos en un dominio reducido que gira en torno a datos disponibles sobre la oferta formativa, su personal docente, y la estructura organizativa de la UNED.

Nuestra intención es modelar los estudios ofertados con una cantidad de información suficiente como para realizar a posteriori consultas de interés que alberguen aplicaciones potenciales útiles para la Universidad. Así, los recursos que hemos decidido modelar son los siguientes:

- Programas de Estudios, incluyendo la facultad a la que pertenecen, las asignaturas que los componen, sus páginas web, la rama de conocimientos donde se enmarcan, qué tipo de estudios contemplan (Máster, Grado, etc.) y sus códigos dentro de la Institución.

- Dado que los Programas de Estudios presentan información sobre las asignaturas que los conforman, sería interesante definir información sobre éstas: en qué curso del programa se enmarca cada asignatura; si una asignatura es de duración anual o cuatrimestral y en ese caso qué cuatrimestre la alberga; cuántos créditos otorga en su plan de estudios; su página web; si es de carácter obligatorio, optativo, etc.; código utilizado por la Universidad para la identificación de la asignatura; profesores que la imparten.

- De los profesores, por el momento, modelaremos información concisa: su nombre, su página web, departamento al que pertenecen, y asignaturas que imparten.

- También almacenaremos información básica sobre departamentos y facultades: nombre, página web, código identificativo dentro de la Institución.

Toda esta información será fácilmente ampliable con nuevos campos y nuevas instancias, pudiendo crecer incrementalmente con datos de varias fuentes (internas o externas a la institución). Así, por ejemplo, con otros fines ajenos pero complementarios al de este trabajo, se puede enriquecer la información de las asignaturas con recursos multimedia o DBpedia relacionados, o la de los profesores con trabajos de investigación y publicaciones realizadas.

3.1.2. Presentación de los datos disponibles, su formato y obtención

Con el objetivo de extraer la información presentada en el apartado anterior, se han realizado una serie de consultas al Portal Estadístico de la Universidad[1], disponiendo los datos de manera estructurada en diversos ficheros CSV (Comma Separated Values).

Aunque para la realización de este trabajo la obtención de datos se ha realizado a través de consultas variadas, descargas y disposición en el formato y los ficheros citados, si este mismo trabajo se desarrollara institucionalmente, la implementación se debería concebir como un proceso de matching desde las bases de datos relacionales de la Universidad hacia el *datastore* (tal y como figura en [46, 18].

Los ficheros descargados y los datos contenidos por cada uno de éstos son los siguientes:

- Facultades.csv: identificador (suele darse por la unión de "fac" y el código identificativo de ésta), código identificativo, nombre de la facultad, página web, identificador de departamento asignado a esta facultad.
- Departamentos.csv: identificador (suele darse por la unión de "dpto" y el código identificativo de éste), código identificativo, nombre del departamento, página web, identificador asociado a profesor adscrito a éste.
- Profesores.csv: identificador del profesor, apellidos y nombre, apellidos, nombre, página web, identificador de asignatura que imparte.
- Grados.csv: identificador de asignatura, código de asignatura, nombre, página web, tipo de asignatura, créditos, duración, semestre, curso, identificador del programa al que pertenece.
- Másteres.csv: identificador de asignatura, código de asignatura, nombre, página web, créditos, tipo de asignatura, especialidad, identificador del programa al que pertenece.
- Acceso.csv: identificador de asignatura, código de asignatura, nombre, página web, identificador del programa al que pertenece.
- Programas.csv: identificador de programa, código de programa, nombre, tipo de estudios, página web, rama de conocimiento, facultad a la que está adscrito.

Como información adicional que acompaña a los ficheros se tiene que:

- Cada fichero puede tener varias líneas para el mismo identificador, tantas como recursos diferentes tenga para un determinado campo.
- Un profesor puede impartir varias asignaturas.
- Un profesor pertenece a un único departamento.
- Un departamento está asignado a una única facultad o escuela.
- Los códigos identificadores de facultad, departamento, asignatura y programa de estudios son únicos.

3.1.3. Diseño de la estructura del *datastore*

Los datos que vamos a modelar han sido obtenidos mediante diversas consultas y vistas que tienen su fuente en la base de datos interna de la UNED. Dicha base de datos tiene una estructura esquemática propia con sus tablas, relaciones, dominios, etc.

[1] `https://app.uned.es/evacal`

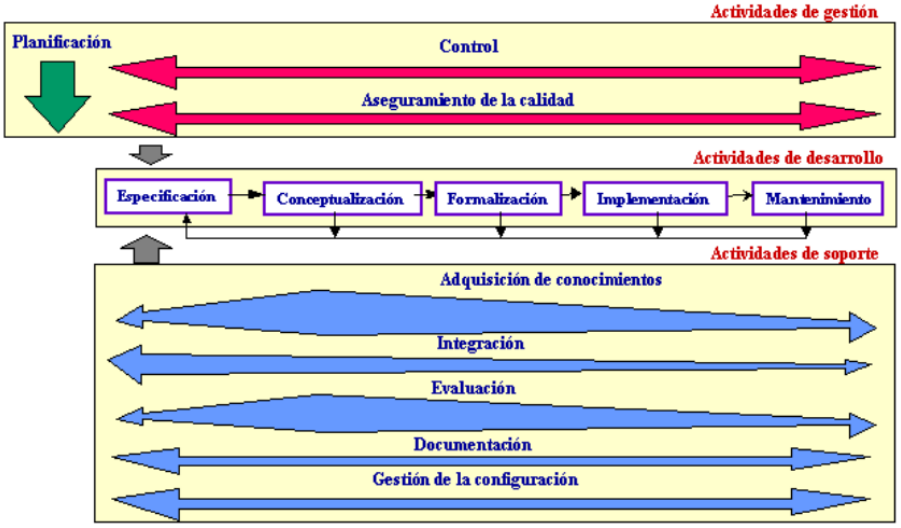


Figura 3.1: Ciclo de actividades en Methontology

La creación del *datastore* puede ser tan rápida y sencilla como se desee: podemos realizar un volcado directo de la base de datos hacia el *datastore* respetando y utilizando las relaciones ya existentes, o podemos diseñar una ontología que nos permita estudiar y enriquecer los conceptos, relaciones y propiedades propios de nuestro *datastore*, estableciendo nuevas relaciones y propiedades (que, por ejemplo, permitan realizar consultas de una forma más directa o realizar inferencias sobre el conjunto de datos).

En este trabajo, para dar forma a la estructura de datos hemos diseñado una pequeña ontología que permitirá que los datos y su estructura puedan crecer o ser modificados fácilmente. Existen diversas metodologías para el diseño y creación de ontologías que modelen un determinado área de conocimiento. En [63] se presenta una breve comparativa entre distintas metodologías, tales como Diligent [49], TOVE [64], Methontology [25], o NeOn [61]. En nuestro caso, al tratarse de un dominio pequeño que se quiere modelar desde cero utilizaremos la metodología Methontology.

Methontology es una metodología desarrollada en el Laboratorio de Inteligencia Artificial de la Universidad Politécnica de Madrid, la cual propone conceptualizar las ontologías utilizando representaciones intermedias tabulares y gráficas que modelen el conocimiento, proporcionando una guía de desarrollo de ontologías a través de diversas actividades (ver figura 3.1). Dicho ciclo de actividades se basa en el proceso de desarrollo de software propuesto por la organización IEEE[2], además de en otras metodologías de ingeniería de conocimientos.

Para asegurar la consistencia y completitud de la ontología, en [19] se propone llevar a cabo una serie de tareas durante la actividad de conceptualización (en la cual nos vamos a centrar en este apartado). Dichas tareas vienen representadas en la figura 3.2 y especificadas, para nuestro caso, a continuación.

En [19] se muestra el proceso de construcción de una ontología legal, haciendo uso de la citada metodología.

TAREA 1: construir el glosario de términos

Durante la primera tarea se deberán identificar todos los términos importantes del dominio (conceptos, atributos, relaciones entre conceptos e instancias); en nuestro caso, dichos términos se reflejan en la tabla 3.1 (a excepción de las instancias, de las que sólo se reflejan unas pocas

[2] `http://standards.ieee.org/develop/index.html?utm_source=mm_link&utm_campaign=faw&utm_medium=std&utm_term=develop%20standards%2C%20find%20working%20group`

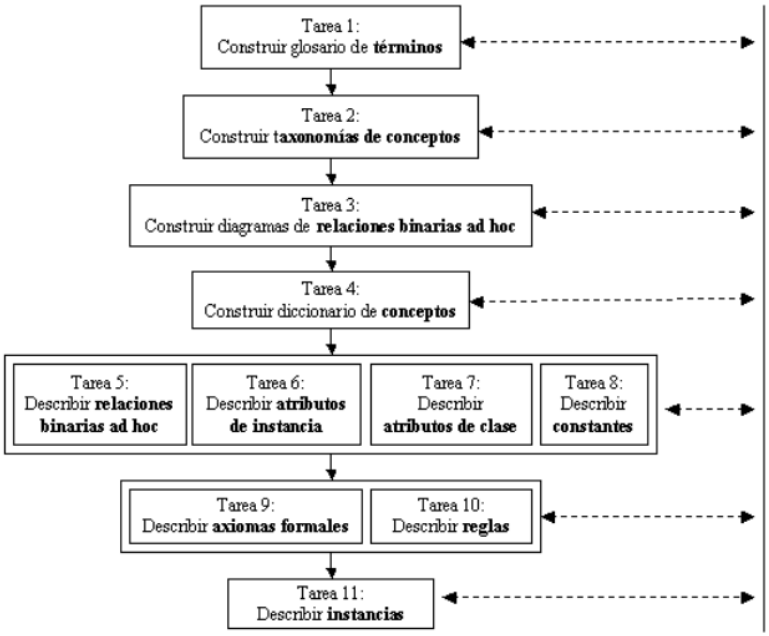


Figura 3.2: Tareas para llevar a cabo la conceptualización en Methontology

por cuestiones de espacio) y se describen brevemente a continuación:

- Facultad: institución docente donde se imparten estudios superiores especializados en alguna materia o rama de conocimiento.
- Departamento: sección en que está dividida una institución u organización, en este caso una facultad.
- Profesor: miembro docente que tiene entre sus cometidos la investigación y la enseñanza de algunas materias.
- Equipo Docente: conjunto de profesores que imparten una misma asignatura.
- Asignatura: materia que forma parte de un programa de estudios.
- Estudios o Programa de Estudios: conjunto de materias que forman un determinado plan de estudios. Puede tratarse de un Grado, un Máster o un Curso determinado. Un Grado es una titulación de educación superior en una o varias disciplinas, orientada a la adquisición de unas competencias necesarias para ejercer una determinada profesión. Un Máster es un curso de especialización en una determinada materia. En este trabajo, el Curso que modelamos es el que presenta las materias que permiten el acceso a la universidad para personas mayores de 25 y 45 años.
- Contiene: relación por la que una facultad contiene uno o varios departamentos.
- Forma parte de: relación inversa a “contiene” por la que un departamento pertenece a una facultad.
- Personal: relación por la que un departamento dispone de uno o varios profesores.
- Es personal de: relación inversa a “personal” que establece que un profesor pertenece a un departamento.

Nombre	**Sinónimos**	**Acrónimos**	**Tipo**
facultad			Concepto
departamento			Concepto
profesor			Concepto
equipo docente			Concepto
asignatura			Concepto
estudios	programa de estudios		Concepto
grado			Concepto
máster			Concepto
acceso			Concepto
contiene			Relación
forma parte de			Relación
personal			Relación
es personal de			Relación
miembro			Relación
es miembro de			Relación
imparte			Relación
es impartida			Relación
adscrito			Relación
tiene adscrito			Relación
pertenece			Relación
tiene asignado			Relación
se enmarca			Relación
contextualiza			Relación
tipo de estudios			Atributo de clase
cuatrimestre			Atributo de instancia
créditos	ECTS		Atributo de instancia
nombre			Atributo de instancia
página web	sitio web		Atributo de instancia
código			Atributo de instancia
descripción			Atributo de instancia
especialidad			Atributo de instancia
curso			Atributo de instancia
duración			Atributo de instancia
modalidad			Atributo de instancia
rama de conocimiento			Atributo de instancia
dpto0804			Instancia
prog2806			Instancia
fac06			Instancia

Tabla 3.1: Glosario de términos

- Miembro: relación por la que un Equipo Docente cuenta con uno o varios profesores como miembros.
- Es miembro de: relación inversa a “miembro”, por la que un profesor pertenece a un equipo docente.
- Imparte: relación que establece que un profesor enseña una asignatura.
- Es impartida: relación inversa a “imparte”, por la que una asignatura es enseñada por un profesor.
- Adscrito: relación que dispone que un Equipo Docente tiene asignada una asignatura.
- Tiene adscrito: relación inversa a “adscrito”, por la que una asignatura tiene asignado un equipo docente.
- Pertenece: relación que indica que un determinado plan de estudios está asignado a una determinada facultad.
- Tiene asignado: relación inversa a “pertenece”, por la que una facultad tiene uno o varios planes de estudios asignados.
- Se enmarca: relación que indica que una asignatura se encuadra en un determinado plan de estudios.
- Contextualiza: relación inversa a “se enmarca”, que establece que un plan de estudios viene determinado por varias asignaturas.
- Tipo de estudios: pueden ser estudios de Grado, de Máster o de Acceso.
- Cuatrimestre: en el que tiene lugar una asignatura dentro de un curso en un programa.
- Créditos: se presentan como unidades de valor de los estudios universitarios para el Espacio Europeo de Educación Superior. Cada crédito equivale a unas 25 ó 30 horas lectivas.
- Nombre: actúa como identificador de entidades, en nuestro caso de facultad, departamento, profesor, asignatura y estudios.
- Página web o sitio web: localización de determinada información dentro de la WWW. Las instancias de facultad, departamento, profesor, asignatura y estudios tendrán su propia página web.
- Código: identifica de manera univoca una entidad.
- Descripción: de un recurso mediante un breve texto.
- Especialidad: se puede optar por diferentes especialidades para estudiar un determinado Máster.
- Curso: en el que se imparte una asignatura dentro de un programa. Puede haber hasta 4 (como es el caso de los Grados).
- Duración: de una asignatura. Puede ser anual o semestral.
- Modalidad: indica si una asignatura es optativa, obligatoria, forma parte de un programa de prácticas, es de formación básica, etc.
- Rama de conocimiento: grandes campos del saber cuya seña de identidad es un conjunto de materias que son la esencia de cada rama.

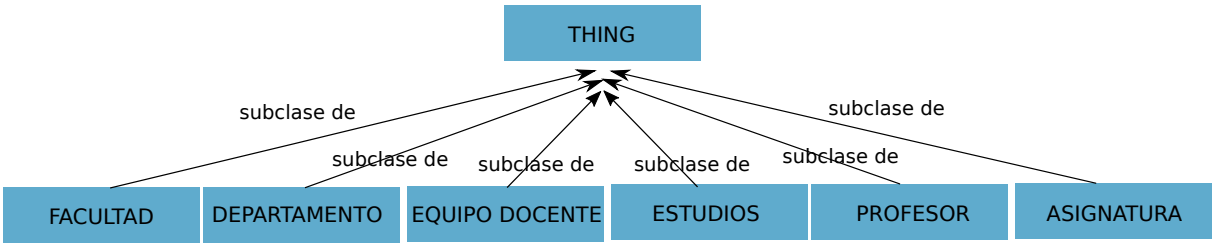


Figura 3.3: Taxonomía de conceptos

TAREA 2: construir taxonomías de conceptos

Tras la definición del glosario de términos se debe definir la jerarquía que gobierna los conceptos construyendo su taxonomía. En nuestro pequeño modelo, por el momento, no existen conceptos que sean subclase de otros conceptos, con lo que estarán todos al mismo nivel, presentando la taxonomía más sencilla posible (figura 3.3).

TAREA 3: construir un diagrama de relaciones binarias ad hoc

La tercera tarea se propone el establecimiento de relaciones ad hoc existentes entre conceptos de la misma o de distintas taxonomías de conceptos. En la figura 3.4 se pueden apreciar las relaciones binarias que se han detectado para nuestro dominio.

TAREA 4: construir el diccionario de conceptos

El siguiente paso consiste en identificar para cada concepto, las propiedades, relaciones e instancias que lo describen, para reflejarlo en un diccionario de conceptos (se especificarán únicamente las relaciones que tienen su origen en el concepto en cuestión).

La tabla 3.2 presenta el diccionario de conceptos definido para nuestro dominio. En la columna de instancias sólo se han identificado unas pocas ya que, por lo general y dependiendo del concepto, puede haber entre decenas y miles.

TAREA 5: describir en detalle las relaciones binarias

Esta tarea tiene como objetivo describir en detalle las relaciones binarias identificadas durante la tarea 3. En la tabla 3.3 se pueden consultar las relaciones que se han identificado en nuestra ontología.

TAREA 6: describir en detalle los atributos de instancias

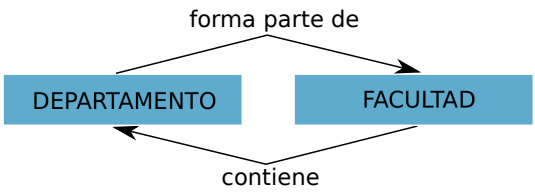
En esta tarea se deben describir, mediante una tabla, los atributos de instancia incluidos en la tarea 1. Los atributos de instancia pueden contener valores distintos para cada instancia de un mismo concepto.

Aquellos detectados para nuestro caso han sido presentados en la tabla 3.4.

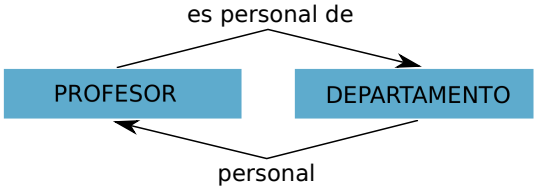
TAREA 7: describir en detalle los atributos de clase

Tras la descripción de los atributos de instancia, se pasa a describir los atributos de clase que fueron identificados en el glosario de términos. A diferencia de los atributos de instancia, los atributos de clase toman el mismo valor para cada una de las instancias de un mismo concepto.

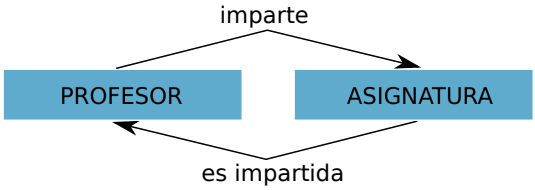
No hay atributos de clase definidos en este trabajo.



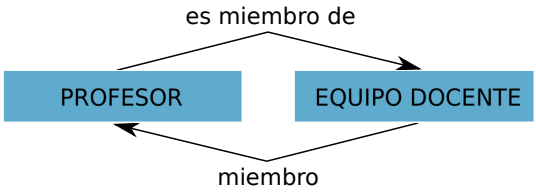
(a) Relación departamento-facultad



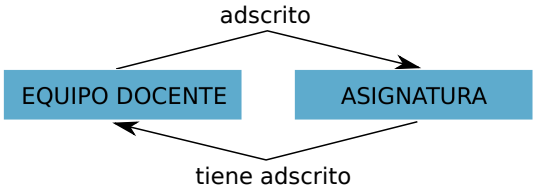
(b) Relación profesor-departamento



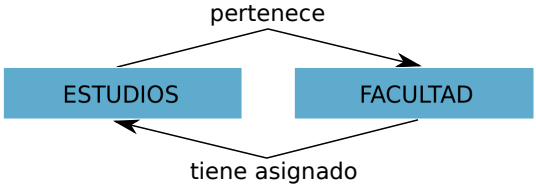
(c) Relación profesor-asignatura



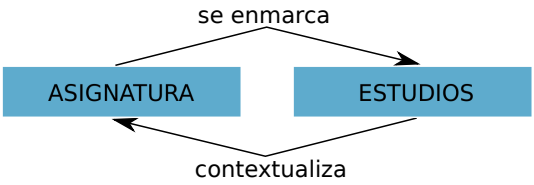
(d) Relación profesor-equipo docente



(e) Relación equipo docente-asignatura



(f) Relación estudios-facultad



(g) Relación asignatura-estudios

Figura 3.4: Diagramas de relaciones binarias

Concepto	**Instancias**	**Atr. de clase**	**Atr. de instancia**	**Relaciones**
facultad	fac01 fac02 fac03 fac04		nombreFacul descripcionFacul paginaWebFacul codigoFacul	contiene tieneAsignado
departamento	dpto0101 dpto0103 dpto0104 dpto0107		nombreDep descripcionDep paginaWebDep codigoDep	formaParteDe personal
profesor	prof00010 prof00020 prof00030		nombreProf paginaWebProf	esPersonalDe esMiembroDe imparte
equipo docente	ED_asig00001028 ED_asig00001092			miembro adscrito
asignatura	asig00001028 asig00001092 asig7002201- asig70022109 asig64011030 asig64022074 asig70013056 asig2440257- asig24400381 asig70022032		cuatrimestre descripcionAsig creditos duracion paginaWebAsig especialidad nombreAsig codigoAsig curso modalidad	esImpartida tieneAdscrito seEnmarca
estudios	prog7102 prog6701 prog6103 prog6603 prog7101		tipoDeEstudios nombreEst paginaWebEst codigoEst ramaDeConocimiento	pertenece contextualiza subclase

Tabla 3.2: Diccionario de conceptos

Nombre de la Relación	**Concepto Origen**	**Cardinalidad Máxima**	**Concepto Destino**	**Relación Inversa**
contiene	Facultad	N	Departamento	formaParteDe
formaParteDe	Departamento	1	Facultad	contiene
personal	Departamento	N	Profesor	esPersonalDe
esPersonal de	Profesor	1	Departamento	personal
miembro	Equipo Docente	N	Profesor	esMiembroDe
esMiembroDe	Profesor	N	Equipo Docente	miembro
imparte	Profesor	N	Asignatura	esImpartida
esImpartida	Asignatura	N	Profesor	imparte
adscrito	Equipo Docente	N	Asignatura	tieneAdscrito
tieneAdscrito	Asignatura	1	Equipo Docente	adscrito
pertenece	Estudios	1	Facultad	tieneAsignado
tieneAsignado	Facultad	N	Estudios	pertenece
seEnmarca	Asignatura	N	Estudios	contextualiza
contextualiza	Estudios	N	Asignatura	seEnmarca

Tabla 3.3: Relaciones binarias

Nombre del Atributo de Instancia	**Concepto**	**Tipo de Valor**	**Rango de Valores**	**Cardinalidad**
cuatrimestre	Asignatura	Entero	1..2	(0,1)
creditos	Asignatura	Decimal	0..N	(1,1)
nombreAsig	Asignatura	Cadena	–	(1,1)
nombreProf	Profesor	Cadena	–	(1,1)
nombreFacul	Facultad	Cadena	–	(1,1)
nombreDep	Departamento	Cadena	–	(1,1)
nombreEst	Estudios	Cadena	–	(1,1)
descripcionAsig	Asignatura	Cadena	–	(0,1)
descripcionFacul	Facultad	Cadena	–	(0,1)
descripcionDep	Departamento	Cadena	–	(0,1)
paginaWebAsig	Asignatura	Cadena	–	(0,1)
paginaWebProf	Profesor	Cadena	–	(0,1)
paginaWebFacul	Facultad	Cadena	–	(0,1)
paginaWebDep	Departamento	Cadena	–	(0,1)
paginaWebEst	Estudios	Cadena	–	(0,1)
codigoAsig	Asignatura	Cadena	–	(1,1)
codigoFacul	Facultad	Cadena	–	(1,1)
codigoDep	Departamento	Cadena	–	(1,1)
codigoEst	Estudios	Cadena	–	(1,1)
especialidad	Asignatura	Cadena	–	(0,1)
curso	Asignatura	Entero	1..4	(0,1)
duracion	Asignatura	Cadena	{anual,semestral}	(0,1)
modalidad	Asignatura	Cadena	{formación básica, obligatorias, optativas, prácticas, trabajo final obligatorio, contenidos, practicum, trabajo de investigación}	(1,1)
tipoDeEstudios	Estudios	Cadena	grado, master, acceso	(1,1)
ramaDeConocimiento	Estudios	Cadena	{Artes y Humanidades, Ciencias, Ciencias de la Salud, Ciencias Sociales y Jurídicas, Ingeniería y Arquitectura}	(0,1)

Tabla 3.4: Atributos de instancia

Nombre de Instancia	Nombre de Concepto	Atributo	Valores
dpto0103	Departamento	nombreDep	QUÍMICA ORGÁNICA Y BIO-ORGÁNICA
dpto0804	Departamento	nombreDep	INGENIERÍA ELÉCTRICA, ELECTRÓNICA Y DE CONTROL
prog2806	Estudios	ramaDeConocimiento	Ingeniería y Arquitectura
prog2806	Estudios	nombreEst	MASTER UNIVERSITARIO EN INGENIERÍA INDUSTRIAL
fac06	Facultad	nombreFacul	FACULTAD DE DERECHO
asig21151075-prog2151	Asignatura	creditos	6
asig21152256-prog2152	Asignatura	especialidad	ESPECIALIDAD: GEOMETRÍA Y TOPOLOGÍA
asig21152256-prog2152	Asignatura	creditos	7.5
asig21152256-prog2152	Asignatura	nombreAsig	TEORÍA DE LA MEDIDA

Tabla 3.5: Instancias

TAREA 8: describir en detalle las constantes

Cada una de las constantes que hemos identificado durante la tarea 1 debe ser descrita en detalle y mediante una tabla en la tarea 8.

No hay constantes definidas en este trabajo.

TAREA 9: definir los axiomas formales

Durante esta tarea se deben identificar los axiomas formales necesarios en la ontología y describirlos en una tabla de manera precisa.

No hay axiomas a definir en este trabajo.

TAREA 10: definir las reglas

Al igual que ocurría en la tarea 9 con los axiomas, se debe identificar qué reglas se necesitan y posteriormente describirlas en una tabla.

No hay reglas a definir en este trabajo.

TAREA 11: describir instancias

Una vez creado el modelo conceptual de la ontología se pueden definir las instacias para cada concepto. En nuestro caso existen decenas de miles de instancias, por lo que en la tabla 3.5 sólo se presentan unas pocas a modo de ejemplo.

3.1.4. Creación de la estructura del *datastore*

Como estructura de nuestro *datastore* vamos a diseñar una pequeña ontología utilizando el editor Protégé [45] (en su versión 5.2.0), dado que ofrece una arquitectura *plug-in* cómoda e intuitiva, la cual, ofrece visualmente un entorno bastante amigable en la construcción de ontologías tan sencillas o complejas como se deseen.

Protégé es un editor de ontologías libre y de código abierto para la construcción de sistemas inteligentes, desarrollado en la Escuela de Medicina de la Universidad de Stanford y que cuenta con el apoyo activo de una gran comunidad de usuarios y desarrolladores.

A través de una interfaz de usuario completamente personalizable se pueden crear y editar ontologías, y a través de las herramientas de visualización que proporciona, permite la navegación interactiva por las relaciones de la ontología. También ofrece apoyo al desarrollador, localizando inconsistencias y dando una explicación avanzada de éstas. Además, Protégé incorpora razonadores como Pellet, FaCT ++ o HerMiT, los cuales hacen uso de los axiomas de la ontología para extraer nueva información a través de inferencias.

El diseño ontológico con Protégé se torna sencillo cuando se trata de un pequeño dominio como el nuestro. Existe bastante documentación en la Web sobre la utilización de esta herra-

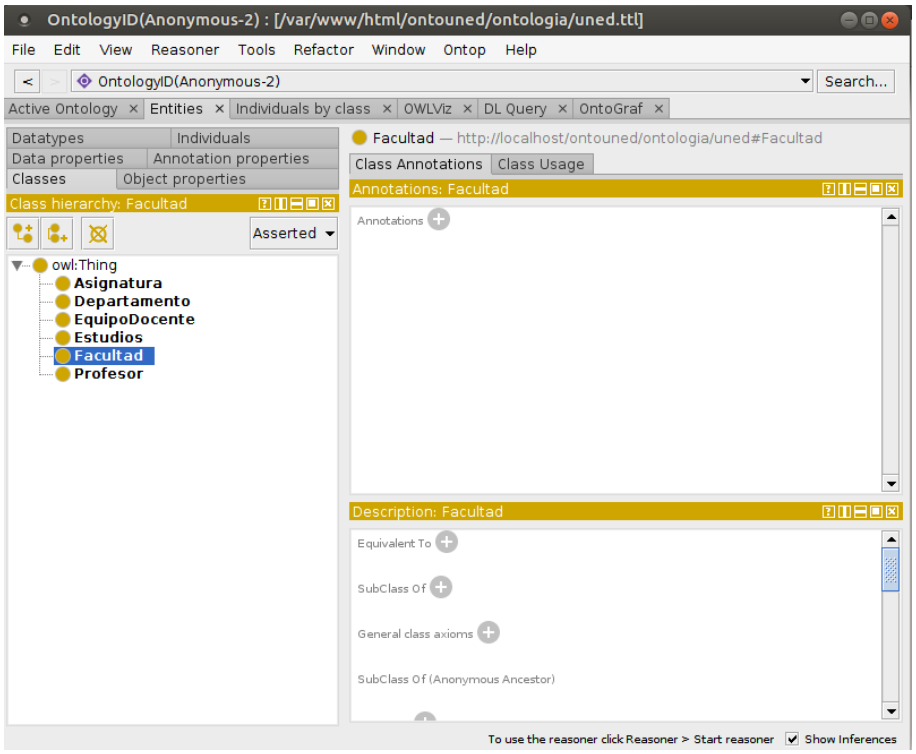


Figura 3.5: Interfaz de exploración y definición de clases en Protégé

mienta, donde se debe hacer especial referencia a la documentación de usuario que encontramos en Protégé Wiki[3].

Disponiendo del diccionario de conceptos diseñado con Methontology que presentabamos en la tabla 3.2, podemos definir de una forma fácil e intuitiva las clases con Protégé en la pestaña "Entidades" dentro de la subpestaña "Clases" (ver figura 3.5).

Tras la definición de las clases, en la subpestaña "Propiedades de Objeto" podemos definir las relaciones que presentabamos en detalle mediante la tabla 3.3, pudiendo establecer propiedades de estas relaciones como, por ejemplo, la funcionalidad o la simetría. En nuestro caso, sólo hemos considerado la propiedad "funcional", disponiéndola en aquellas relaciones de cardinalidad máxima 1. En la figura 3.6 se presenta la interfaz de exploración y definición de relaciones en Protégé.

Para terminar de crear la ontología con los términos identificados con Methontology (a excepción de las instancias) nos faltan por definir los atributos. Para ello, dentro de la pestaña "Entidades" y dentro de la subpestaña "Propiedades de Datos" en Protégé (ver figura 3.7), modelaremos la información que identificamos y describimos en detalle mediante la tarea 6 del proceso de conceptualización de Methontology (ver tabla 3.4).

Protégé también permite obtener una visualización gráfica del conjunto o de una parte de las clases y sus relaciones mediante la pestaña "OntoGraf". Podemos recurrir a la figura 3.8 para visualizar el grafo que define nuestro dominio de estudio.

3.1.5. Poblado automático desde fuentes estructuradas

Hasta ahora, hemos estudiado la estructura de nuestro *datastore* y la hemos implementado haciendo uso de la herramienta Protégé. Sin embargo, llegado este momento, necesitamos una herramienta que permita el poblado automático de nuestra pequeña ontología con las decenas de miles de datos que tenemos disponibles en ficheros CSV. Para ello, hacemos uso del *framework*

[3] `https://protegewiki.stanford.edu/wiki/Main_Page`

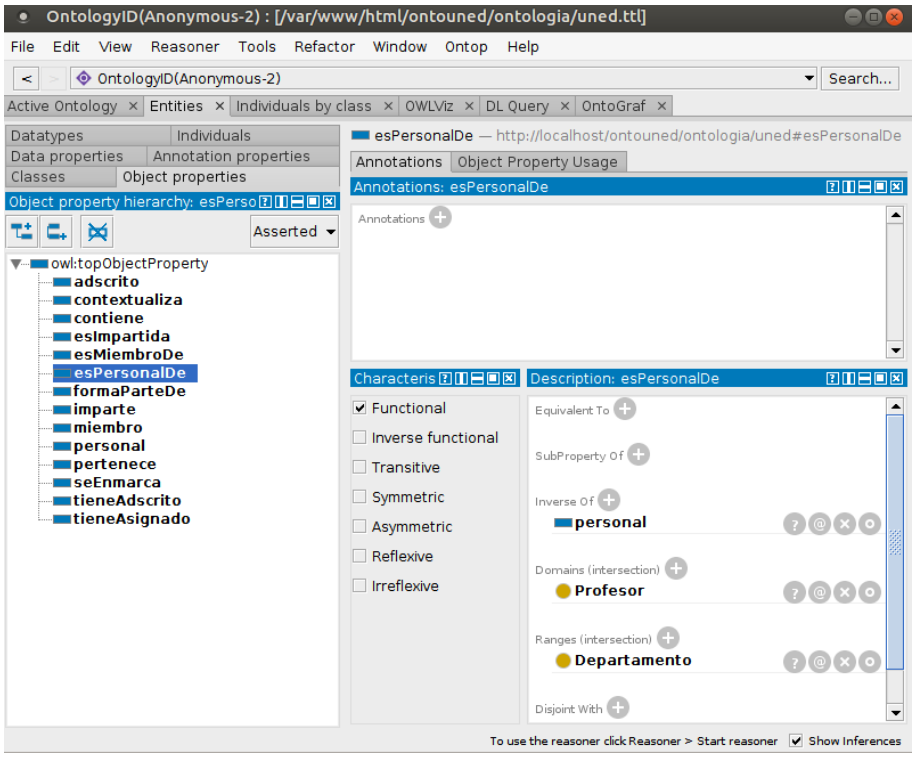


Figura 3.6: Interfaz de exploración y definición de propiedades de objeto en Protégé

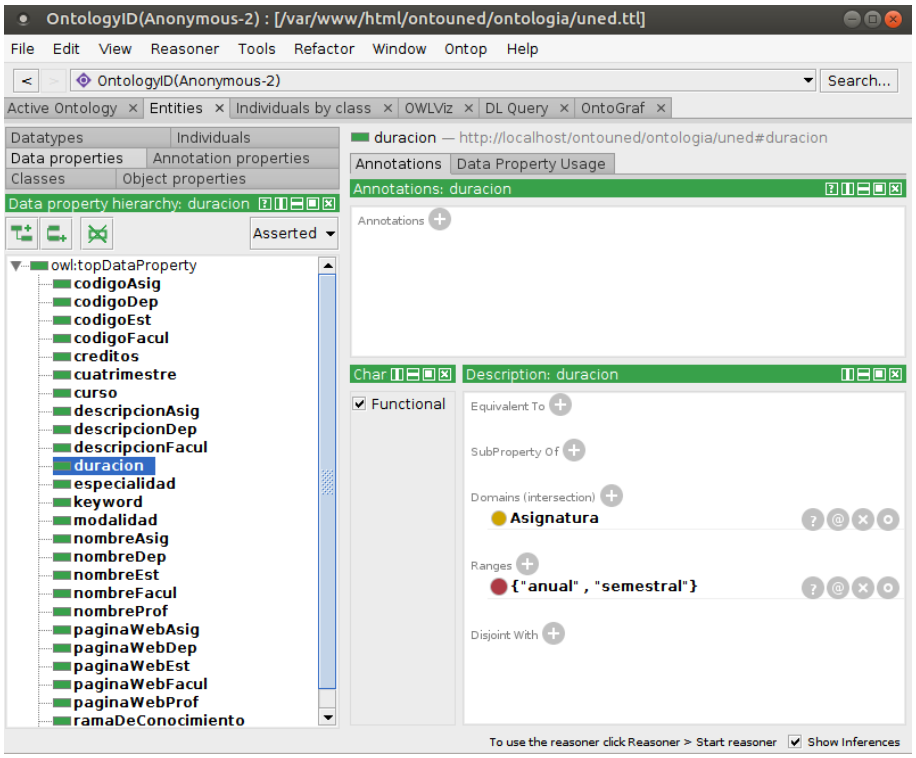


Figura 3.7: Interfaz de exploración y definición de propiedades de datos en Protégé

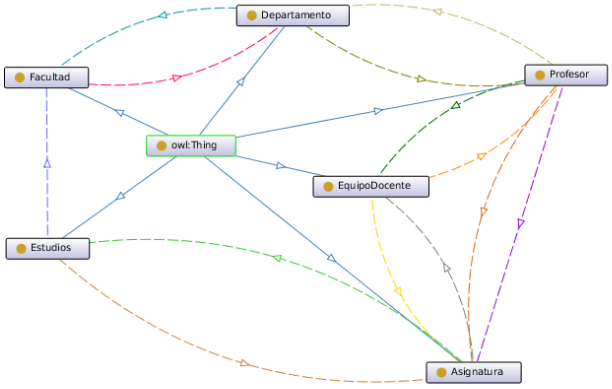


Figura 3.8: Grafo de la ontología

Apache Jena, dado que es fácilmente integrable en una interfaz de desarrollo, que además permita interactuar con otras librerías capaces de explorar y explotar ficheros CSV, bases de datos, *web crawlers*, etc.

Apache Jena

Haciendo referencia a la información disponible en su sitio web[4], Jena es un *framework* de código abierto creado para trabajar en el marco de las tecnologías de la Web Semántica y Datos Enlazados. Dispone de una extensa colección de librerías Java para ayudar a los desarrolladores a trabajar con RDF, RDFS, RDFa, OWL y SPARQL, de acuerdo con las recomendaciones W3C. Dicho *framework* incluye un motor de inferencias basado en reglas para mejorar el razonamiento para ontologías OWL y RDFS. Además incluye una gran variedad de estrategias de almacenamiento de tripletas RDF, tanto en memoria como en disco.

Fue desarrollado por investigadores de Laboratorios HP en el año 2000, como un software de código abierto diseñado para trabajar con aplicaciones en la Web Semántica a través de Java. En noviembre de 2010 Jena pasa a formar parte de la Fundación de Software Apache, donde continúa su desarrollo hasta la actualidad.

El framework de Jena incluye:

- Un API[5] para leer, procesar y escribir datos RDF en los formatos XML, N-Triples y Turtle.
- Un API para el manejo de ontologías OWL y RDFS.
- Un motor de inferencias basado en reglas con capacidad de razonamiento con fuentes de datos RDF y OWL.
- Almacenes de tripletas RDF con gran capacidad y eficiencia de almacenamiento en disco.
- Un motor de consultas compatible con las últimas especificaciones SPARQL.
- Servidores para la publicación de datos RDF para su uso por otras aplicaciones, mediante la utilización de gran variedad de protocolos.

[4] `https://jena.apache.org/index.html`
[5] Interfaz de Programación de Aplicaciones

El API de Jena se puede integrar fácilmente con un IDE[6] como Eclipse[7] o Netbeans[8]; existe una amplia documentación del manejo de las librerías en la página de tutoriales[9] del sitio web de Apache Jena.

Poblado automático con Jena desde fuentes de datos CSV

Dado que disponemos de gran cantidad de datos estructurados en varios ficheros CSV, recorreremos todos estos datos creando los individuos de cada clase y asignándoles las propiedades pertinentes.

Para llevar a cabo esta tarea, se ha utilizado el IDE Netbeans con las APIs JavaCSV[10] para recorrer los ficheros CSV, y Jena para recorrer y poblar la ontología creada con Protégé. En el siguiente fragmento de código 3.1 se muestra el proceso realizado para el poblado de la ontología con el fichero "facultad.csv". A modo de resumen, el código realiza los siguientes pasos:

1. Carga la ontología creada con Protége mediante la sentencia "OntModel...".
2. Carga los recursos y propiedades que se van a utilizar: facultad, departamento, nombre, web, descripción, código, etc.
3. Recorre el fichero CSV línea por línea:
   a) Si el recurso facultad al que estamos accediendo ya había sido creado, se carga y se le asignan las propiedades, accediendo a estas en el fichero CSV mediante la orden reader(columna).
   b) Si el recurso facultad al que estamos accediendo no existe, se crea y se le asignan las propiedades, accediendo a estas en el fichero CSV mediante la orden reader(columna).
4. Por último, se almacena el modelo ya poblado.

Código 3.1: Poblado del datastore con datos del fichero facultad.csv

```
1  OntModel model = ModelFactory.createOntologyModel(OntModelSpec.OWL_MEM);
2  model.read("uned.ttl", "Turtle");
3
4  //Obteniendo los recursos
5  OntClass departamentoR = model.getOntClass(ontuned + "Departamento");
6  OntClass facultadR = model.getOntClass(ontuned + "Facultad");
7
8  //Creando las propiedades
9  ObjectProperty contieneP = model.getObjectProperty(ontuned + "contiene")
       ;
10 ObjectProperty formaParteDeP = model.getObjectProperty(ontuned + "
       formaParteDe");
11
12 DatatypeProperty nombreFaculP = model.getDatatypeProperty(ontuned + "
       nombreFacul");
13 DatatypeProperty descripcionFaculP = model.getDatatypeProperty(ontuned +
        "descripcionFacul");
14 DatatypeProperty webFaculP = model.getDatatypeProperty(ontuned + "
       paginaWebFacul");
```

[6] Entorno de Desarrollo Integrado
[7] https://eclipse.org/
[8] https://netbeans.org/
[9] https://jena.apache.org/tutorials/index.html
[10] https://www.csvreader.com/java_csv.php

```
15 DatatypeProperty codigoFaculP = model.getDatatypeProperty(ontuned + "
       codigoFacul");
16
17 //Poblando con el fichero de facultades.csv
18 CsvReader reader = null;
19
20 try {
21     InputStreamReader inputStreamReader = new InputStreamReader(new
           FileInputStream("facultades.csv"), "UTF8");
22
23     //Creo los individuos de facultad y realizo su poblado
24     reader = new CsvReader(inputStreamReader);
25     reader.setDelimiter(';');
26     reader.readHeaders();
27
28     while (reader.readRecord()) {
29         if (model.containsResource(model.getResource(ontuned + reader.
               get(0)))) {
30             Individual ind = model.getIndividual(ontuned + reader.get(0))
                   ;
31             ind.setPropertyValue(codigoFaculP, model.createLiteral(reader
                   .get(1)));
32             ind.setPropertyValue(nombreFaculP, model.createLiteral(reader
                   .get(2)));
33             ind.setPropertyValue(webFaculP, model.createLiteral(reader.
                   get(3)));
34             ind.setPropertyValue(descripcionFaculP, model.createLiteral("
                   Facultad o Escuela Tecnica de la UNED"));
35
36             Resource resdep = model.getResource(ontuned + reader.get(4));
37             Individual ind2 = null;
38             if (model.containsResource(resdep)) {
39                 ind2 = model.getIndividual(ontuned + reader.get(4));
40             } else {
41                 ind2 = model.createIndividual(ontuned + reader.get(4),
                       departamentoR);
42             }
43             ind2.setPropertyValue(formaParteDeP, ind);
44             ind.addProperty(contieneP, ind2);
45
46          } else {
47
48              Individual ind = model.createIndividual(ontuned + reader.get
                    (0), facultadR);
49              ind.setPropertyValue(codigoFaculP, model.createLiteral(
                    reader.get(1)));
50              ind.setPropertyValue(nombreFaculP, model.createLiteral(
                    reader.get(2)));
51              ind.setPropertyValue(webFaculP, model.createLiteral(reader.
                    get(3)));
52              ind.setPropertyValue(descripcionFaculP, model.createLiteral
                    ("Facultad o Escuela Tecnica de la UNED"));
53
54               Resource resdep = model.getResource(ontuned + reader.get(4)
                     );
55              Individual ind2 = null;
56              if (model.containsResource(resdep)) {
57                  ind2 = model.getIndividual(ontuned + reader.get(4));
```

```
58                } else {
59                     ind2 = model.createIndividual(ontuned + reader.get(4),
                           departamentoR);
60                }
61                ind2.setPropertyValue(formaParteDeP, ind);
62                ind.addProperty(contieneP, ind2);
63           }
64      }
65  } catch (Exception e) {
66      e.printStackTrace();
67  } finally {
68      reader.close();
69  }
70
71  BufferedWriter guardarFlujo = new BufferedWriter(new OutputStreamWriter(
       new FileOutputStream("uned.ttl"), "UTF8"));
72  model.write(guardarFlujo, "Turtle");
73  guardarFlujo.close();
```

3.2. Enriquecimiento del *datastore* con datos extraídos desde fuentes desestructuradas mediante técnicas de Minería de Textos

Una vez se ha diseñado y poblado automáticamente la ontología con los datos obtenidos desde fuentes estructuradas de la Universidad, queremos incrementar el poblado de cada asignatura con el campo *keywords*, el cual contendrá términos importantes con capacidad de definir el contenido de cada una de estas asignaturas. Para conseguirlo se va a utilizar la página web de Contenido de las Guías de las asignaturas de la Universidad. Es en este punto donde una simple exploración de los datos del *datastore* RDF inicial, nos proporcionará el acceso a recursos analizables que nos permitirán continuar poblando el *datastore* con nuevos datos obtenidos. Se pretende el diseño de un sistema de recopilación automática, en un único corpus, de todas las páginas de "Contenido" de las más de 3000 guías didácticas, publicadas en la Web, de las asignaturas de la UNED.

El procedimiento de obtención de la información con la que después se incrementará el contenido del *datastore* viene escenificado en la figura 3.9. Dicho proceso consta de los siguientes pasos:

1. Detección de enlaces a las páginas de Contenido y su recuperación automática. Estas páginas se encuentran alojadas en la web de la UNED, cada una con una URL determinada, conocida y contenida en el *datastore* creado hasta ahora, apareciendo como atributo de instancia para cada asignatura. Tras la recuperación de cada enlace mediante consultas sencillas al *datastore* de partida, se descargarán todos los documentos (guías) formando el corpus de textos sobre el que trabajar.

2. Preprocesado de las guías: extracción y limpieza de los términos. Para cada guía se detectarán todos los términos que la forman, y a dichos términos se les asignará un etiquetado morfológico extrayendo, además, su forma canónica. Por otro lado, se eliminarán aquellas palabras que, por sí sólas carezcan de semántica (palabras vacias o *stopwords*): determinantes, conjunciones, etc.

3. Ponderación terminológica: análisis del corpus de guías y estudio comparado de las funciones de pesado. Cada término de cada guía recibirá un valor que simbolizará lo importante que es éste para la representación de la guía a la que pertenece. El objetivo es que los términos con mayor importancia para cada asignatura obtengan los mayores pesos de manera

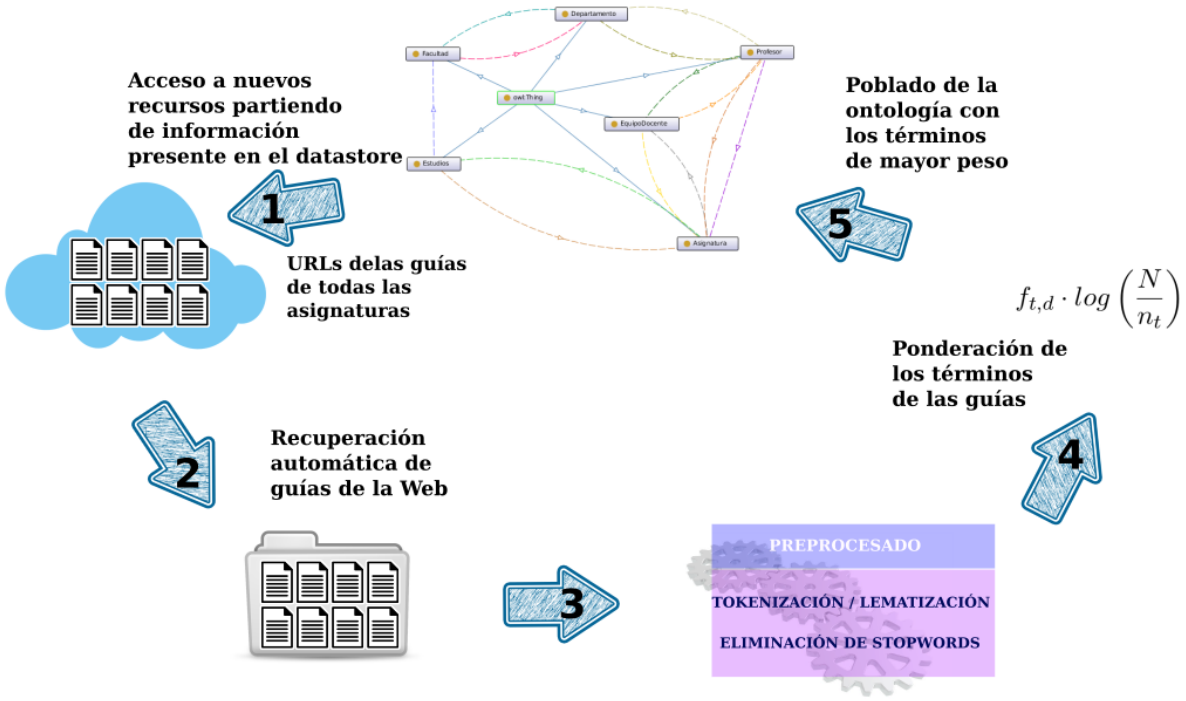


Figura 3.9: Proceso de obtención y poblado del campo *keywords* de la ontología

que, posteriormente, se pueda realizar una extracción controlada de éstos para poblar la ontología.

4. Poblado del *datastore* inicial con los términos extraídos: tras la ponderación y extracción de información terminológica desde las páginas de contenido de las guías de las asignaturas, se completará el campo *keywords* de la ontología con los términos apropiados.

3.2.1. Detección de enlaces a las páginas de Contenido y su recuperación automática

La recuperación de las guías[11] de cada asignatura de la UNED desde sus correspondientes páginas en Internet puede resultar un trabajo muy laborioso si se realiza manualmente, debido a la amplia oferta educativa de esta universidad. Es por esto que dicho proceso se automatizará, con el fin de que esta recuperación se produzca de la manera más rápida y fiable posible.

Entre los datos estructurados con los que se pobló la ontología, se hallan las URL de las direcciones web donde se alojan las guías de las asignaturas. Para acceder a la sección de contenido de una asignatura determinada, habrá que concatenar con la URL de su guía la cadena opcional "*&idContenido=2*" si se trata de una asignatura del Curso de Acceso, o "*&idContenido=5*" si se trata de una asignatura de Grado o de Máster. El procedimiento de obtención de las distintas URLs desde el *datastore* es bastante sencillo y para su puesta en marcha hemos empleado la librería Jena con el código 3.2.

Código 3.2: Recuperación de las web en el *datastore* con Jena

```
74 for (Iterator it = asignaturaR.listInstances(true); it.hasNext();) {
75     Individual ind = (Individual) it.next();
```

[11]Por cuestión de sencillez, a menudo, se hace referencia con la palabra clave "guía" a la sección de Contenido de cada Guía de Estudios, dado que de todas las secciones de esta Guía, la de Contenidos será la única en la que nos centraremos en este trabajo.

```
76      if (ind.isIndividual()&&ind.getPropertyValue(webAsigP)!=null) {
77          String web = ind.getPropertyValue(webAsigP).toString();
78          Individual ind2 = model.getIndividual(ind.getPropertyValue(
                seEnmarcaP).toString());
79          if (ind2.isIndividual()&&ind2.getPropertyValue(tipoDeEstudiosP)
                !=null) {
80              if (ind2.getPropertyValue(tipoDeEstudiosP).toString().equals
                    ("ACCESO")) {
81                  web = web + "&idContenido=2";
82              } else {
83                  web = web + "&idContenido=5";
84              }
85              descargarPagina(web);
86           }
87       }
88  }
```

Dado que disponemos de las direcciones URL donde se alojan las guías de contenido de las asignaturas, podemos descargar el código fuente de sus correspondientes páginas web y *parsear* y extraer de este código los datos etiquetados con HTML que nos interesen. En este trabajo, este procedimiento se ha realizado con la librería jsoup[12] de Java.

Una vez realizado el proceso de descarga y extracción automática de la información contenida en las páginas web de las guías con jsoup, se dispone de un almacen que contiene todos los textos sobre los que se llevarán a cabo los procesos de análisis y extracción terminológica.

3.2.2. Preprocesado de las guías: extracción y limpieza de los términos

Disponemos de los textos que componen la sección de contenidos de la guía de cada asignatura en la Web. Para poder evaluar la representatividad de cada uno de los términos que componen cada guía, se debe llevar a cabo un preprocesado de los textos.

En primer lugar, se debe llevar a cabo una ***tokenización***, proceso mediante el cual se convierten en términos candidatos aquellas secuencias de caracteres que componen el texto. Es decir, el proceso de *tokenización* separa e indexa cada uno de los términos de un documento.

Aunque al hablar de *tokenización* se puede pensar en la división del texto en entidades, donde cada una de las cuales conforman un término como una de las palabras contenidas en un texto (monotérmino), en ocasiones, varias palabras deberían aparecer juntas formando el mismo término (multitérmino), ya que de otro modo su semántica se vería alterada. Ejemplo de ello son los términos “Inteligencia Artificial” y “Alfonso X”.

En los textos muchas palabras pueden aparecer con distintas formas aunque procedan de la misma raíz. Por ejemplo, las palabras “etiquetó”, “etiquetará”, y “etiquetaran”, son distintos tiempos del verbo “etiquetar”. Otro ejemplo lo encontramos con los géneros y los números, como en el caso de “entidad” y “entidades” o de “despistada” y “despistado”. Es necesario que al estudiar las frecuencias con las que aparecen los términos, ya sea en un texto o en una colección de ellos, se contabilicen como iguales todos aquellos que lo sean aunque aparezcan en sus distintas variantes. Tradicionalmente existen dos formas de llevar a cabo este proceso [26]:

1. **Stemming:** consiste en reducir una palabra a su raíz o *stem* recortando su afijo o terminación. Es una aproximación bastante simple que puede fallar con las derivaciones de un término que presenten otra raíz, o con términos diferentes que compartan una idéntica. El *stemming* no tiene en cuenta el contexto en el que aparece la palabra en el texto por

[12] jsoup es una librería de Java diseñada para trabajar con HTML. Proporciona una interfaz de programación muy útil para extraer y manipular datos, utilizando, entre otros, métodos DOM, CSS y jquery. El sitio web oficial de esta librería de código abierto es https://jsoup.org

lo que, los términos "bajó" como forma verbal, "bajo" de instrumento musical, o "bajo" preposición, se presentarán bajo la misma raíz, contabilizando como si del mismo término se tratase. Pero, a pesar de los inconvenientes citados, el *stemming* es una técnica muy utilizada en IR que ha demostrado funcionar bastante bien [35].

Los algoritmos más utilizados para realizar este proceso son, *Porter*[51] para textos en inglés y *Snowball*[52] para textos en castellano.

2. ***Lematización:*** consiste en extraer el lema (o forma canónica) de cada término del documento. Para ello, se tiene en cuenta el contexto en el que éste se envuelve. El proceso radica en realizar un análisis morfológico de las palabras de cada oración. Dicho análisis consiste en determinar la categoría gramatical de cada palabra, lo cual elimina aquellas ambigüedades que el *stemming* era incapaz de manejar, aunque computacionalmente resulta más costoso que éste. Además, el proceso de análisis asigna una etiqueta a cada término, lo que puede ser bastante beneficioso ya que, de esta manera se dispone de información adicional sobre cada término[13]. La asignación de etiquetas a los términos se conoce como etiquetado morfológico, etiquetado *PoS* o *Part-of-Speech Tagging*.

   Éste será el enfoque seguido en este trabajo para llevar a cabo la normalización de los términos. De esta manera, además de obtener la forma canónica de cada término, dispondremos de información adicional; y además de evitar las ambigüedades que podían aparecer con el truncado, se abre la opción de comprobar en trabajos futuros si filtrando por ciertas clases de términos, se puede llegar a mejorar la extracción de aquellos con mayor importancia.

Una vez tenemos los términos identificados y etiquetados se procede a limpiar aquellos que carecen de contenido: caracteres especiales y palabras vacias (***stopwords***). Existen listas de *stopwords* en diferentes idiomas, dentro de las cuales encontramos términos como determinantes, conjunciones, algunos verbos auxiliares, etc. La eliminación de este tipo de palabras es un proceso que se viene utilizando en diferentes técnicas de procesamiento lingüístico y que, normalmente, desemboca en una mejora del rendimiento del sistema [11].

Para realizar el procesamiento del contenido de los textos, se ha hecho uso de la librería de código abierto para C++, **Freeling**[14].

Freeling ha sido desarrollada por El Centro de Investigación TALP de la Universidad Politécnica de Cataluña, el cual realiza el mantenimiento de ésta. Dicha herramienta proporciona distintas funcionalidades en el análisis del lenguaje para gran variedad de idiomas diferentes: inglés, español, portugués, francés, italiano, ruso, alemán, etc. Además, cabe recalcar que en su versión estándar, Freeling está entrenada con noticias de periódico, por lo que es en este tipo de textos donde mejor realiza sus funciones.

Además, Freeling se presenta bajo una arquitectura cliente-servidor [47], lo cual le confiere reusabilidad y eficiencia. Dicha herramienta se compone de los siguientes módulos, algunos de los cuales pueden ser configurados según las necesidades de cada usuario:

- *Tokenización*: divide un texto en una lista de palabras.
- Separación de oraciones: divide el texto en oraciones.
- Detección y normalización de expresiones numéricas.
- Detección y normalización de expresiones en formato de fecha y hora.
- Diccionario morfológico

[13] Según [65], los nombres y los adjetivos suelen ser mejores candidatos como términos importantes, que los verbos y los adverbios.

[14] `http://nlp.lsi.upc.edu/freeling`

- Detección de palabras múltiples: detecta aquellos términos compuestos por varias palabras (multitérminos), y los agrupa en una única entidad, separando dichas palabras por guiones bajos: p. ej. "inteligencia_artificial".
- Detección y clasificación de entidades nombradas: reconocimiento de nombres propios.
- Etiquetado *PoS*: cada palabra es anotada con información morfosintáctica a través de etiquetas propuestas por el grupo EAGLES[15].
- Codificación fonética de una palabra en el estándar *de-facto* SAMPA[16].
- Módulo UKB para la desambiguación lingüística: recibe una frase y clasifica los posibles sentidos que puede tener una palabra en el contexto dado.
- Módulos para el análisis sintáctico de los elementos constituyentes de una oración, teniendo en cuenta las dependencias que guarden entre sí.

Mediante la utilización de Freeling se realizarán los procesos de *tokenización* y *lematización*, llevando a cabo un etiquetado morfológico de los términos.

Freeling permite al usuario realizar la selección de los módulos que va a utilizar. Para nuestro objetivo, es interesante la detección y extracción de multitérminos[17], así como de entidades nombradas. Veamos un ejemplo de refuerce esta afirmación. Disponemos de las guías de las asignaturas "Implicaciones Educativas de la Inteligencia Emocional" y "Técnicas de Inteligencia Artificial en la Ingeniería", las cuales comparten la palabra "inteligencia" entre sus contenidos. Si obviamos la utilización del módulo de detección de palabras múltiples de Freeling, ambas guías estarían relacionadas por el término "inteligencia", perdiendo parte de la semántica que se podría obtener. Sin embargo, si hacemos uso del citado módulo, se detectarán los multitérminos "inteligencia_artificial" e "inteligencia_emocional", lo cual parece más indicado para una representación más adecuada de las guías.

En cuanto a la detección de expresiones numéricas, fechas y horas, será realizada para llevar a cabo su posterior limpieza junto con la de otras *stopwords*, con lo que consecuentemente, la normalización de estos datos es irrelevante.

Tras el preprocesado con Freeling, se lleva a cabo el proceso de limpieza de aquellos términos vacios. Para esto, se han eliminado las *stopwords* haciendo uso de una lista predefinida[18] para palabras vacías del castellano. Además, para realizar una limpieza más sofisticada, se ha recurrido a la utilización de las etiquetas con las que Freeling ha anotado cada término. Así, se han eliminado aquellos términos cuyas etiquetas los clasificaban como adjetivos ordinales, determinantes, pronombres, conjunciones, interjecciones, preposiciones, signos de puntuación, expresiones numéricas, fechas y horas.

3.2.3. Ponderación terminológica: análisis del corpus de guías y estudio comparado de las funciones de pesado

El análisis del corpus nos va a permitir tener una visión global de los términos que éste contiene y cómo se distribuyen entre los documentos (guías) que lo forman. Este análisis nos permitirá comprender la naturaleza de los datos, cuestión importante para entender los resultados finales de este trabajo.

Las guías de las asignaturas, en su sección de contenidos, ofrecen una organización temática de su programa, estando divididas en las "unidades didácticas", "bloques" o "temas" que se

[15] `http://nlp.lsi.upc.edu/freeling-old/doc/tagsets/tagset-es.html`
[16] `https://www.phon.ucl.ac.uk/home/sampa`
[17] En Freeling los términos que forman un multitérmino aparecen unidos por un guión bajo.
[18] `http://www.ranks.nl/stopwords/spanish`

0 %	**5 %**	**10 %**	**15 %**	**20 %**	**25 %**	**30 %**
2.00	18.00	25.00	31.00	38.00	43.00	49.00
35 %	**40 %**	**45 %**	**50 %**	**55 %**	**60 %**	**65 %**
56.00	63.00	70.00	80.00	90.00	105.00	118.45
70 %	**75 %**	**80 %**	**85 %**	**90 %**	**95 %**	**100 %**
140.00	163.75	195.00	231.05	293.00	417.35	27376.00

Tabla 3.6: Cuantiles para la distribución de longitudes de guía en términos

trabajarán en la asignatura. De esta manera, algunas guías se dividen en "bloques temáticos", otras en "temas", otras en "secciones", "apartados", etc. Estas divisiones y su correspondiente desarrollo varía considerablemente de unas guías a otras, dando lugar a una enorme variedad en cuanto a estructura y amplitud de desarrollo[19].

Todas las guías de asignaturas de Máster, Grado y Acceso han sido recopiladas formando un corpus de 3143 documentos. Tras haberse realizado el procesado y limpieza de los textos, los términos contenidos en éstos se encontrarán en su forma canónica, y no encontraremos entre ellos ninguna *stopword*.

El conjunto de guías se compone de un total de 445391 términos repartidos de manera muy heterogénea. Existe una guía con 27376 términos[20], 21 guías de entre, aproximadamente, 1000 y 1500 términos, y 195 guías con menos de 20 términos. Para una mejor observación de esta distribución, en la tabla 3.6 se muestra el corpus dividido en un número de cuantiles determinados por la *regla de Sturges*[21].

Como se observa, existen considerables diferencias en cuanto a la **cantidad de términos por guía**. Sobre todo en el cuantil al 90 % (C_{90}), donde como veremos a continuación, se concentran valores extremos (*outliers*) del conjunto. Analizando el corpus completo, se obtiene que la media de términos por guía es 142, la mediana es 80 y la desviación típica es de 512 términos. Dichos datos son indicativos de una distribución desigual que en sus valores medios se ve afectada afectada por los *outliers* que parece haber en el corpus. Se debe tener en cuenta que en C_{90} (tabla 3.6) hay aproximadamente 20 guías con más de 1000 términos, una de ellas con más de 27000, y varias de entre 300 y 1000 términos, y que el resto del conjunto (un 90 % de los datos) tiene valores por debajo de 300 y un 80 % por debajo de 200. Si truncamos C_{90}, la media recalculada pasa de los 142 términos por guía a 92, la mediana pasa de 80 a 70 términos y la desviación típica pasa de ser 512 a ser de 68 términos términos por guía.

En las gráficas de la figura 3.10 se muestran las frecuencias de las guías por la cantidad de términos que contienen. De los datos obtenidos, se puede observar que existe una gran concentración de guías que contienen, aproximadamente, entre 10 y 90 términos.

Distribución de los términos en el corpus

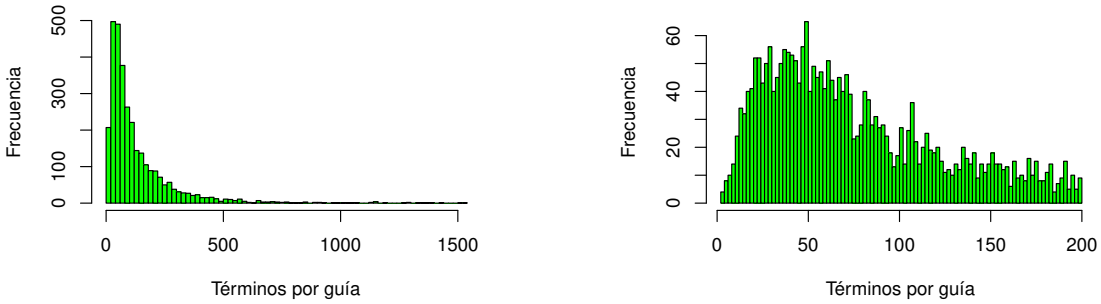
El primer punto a estudiar en este subapartado es la **frecuencia con la que aparecen los términos en el corpus**.

[19] Ejemplos donde se puede apreciar la diferencia en cuanto a estructura y amplitud de desarrollo de las guías:

- HISTORIA MODERNA (Hª DEL ARTE)
- INTRODUCCIÓN A LA PROGRAMACIÓN PARA LA RED
- ACTORES Y COMPORTAMIENTO POLÍTICO
- PREDICCIÓN EN ECONOMÍA

[20] DEMOCRACIA Y DICTADURA EN AMÉRICA LATINA DESDE LA REVOLUCIÓN CUBANA

[21] La regla de Sturges divide un conjunto en un número adecuado k de intervalos. Dicho número viene determinado por $k = 1 + 3.322 \cdot log(n)$.



(a) Histograma de términos por guía

(b) Histograma del rango de 0 a 200 términos

Figura 3.10: Histograma por longitudes de guía

0 %	**4 %**	**8 %**	**12 %**	**16 %**	**20 %**	**24 %**	**28 %**	**32 %**	**36 %**	**40 %**	**44 %**	**48 %**
1	1	1	1	1	1	1	1	1	1	1	1	1
52 %	**56 %**	**60 %**	**64 %**	**68 %**	**72 %**	**76 %**	**80 %**	**84 %**	**88 %**	**92 %**	**96 %**	**100 %**
1	1	2	2	2	3	3	4	6	9	16	42	10855

Tabla 3.7: Cuantiles para la distribución de frecuencias terminológicas en el corpus

El corpus o colección de guías contiene un total de 445391 términos (libres de *stopwords*). Dentro de dicha cantidad, los términos pueden estar repetidos o aparecer de manera única en toda la colección. Realizando un análisis, los términos aparecen en éste con una media de 10.5 repeticiones, una mediana de 1 y una desviación típica de 79.62 términos. De nuevo, hay indicativos de que existen *outliers* en el conjunto de datos. En el histograma de la figura 3.11 se puede observar la distribución de frecuencias de aparición de los términos en el corpus. Además, en la tabla 3.7 se recogen los cuantiles[22]. En dicha tabla, se observa que prácticamente el 85 % de los términos que aparecen en la colección, se presentan con una frecuencia muy escasa en el corpus y que el 15 % restante tiene frecuencias muy variables, que crecen desde las 6 repeticiones en la colección, a miles de repeticiones en ésta.

Dado que existen términos que se repiten con gran frecuencia, e incluso algunos que se repiten con extraordinaria frecuencia, podemos pensar que la causa de este fenómeno podría derivarse de estas dos situaciones:

1. Dada la variedad de tamaños entre las guías, puede haber términos en aquellas guías más largas que superen considerablemente la media normal de apariciones de un término en la colección, dado que en los documentos largos se tiende a repetir frecuentemente los mismos términos[50].

2. Al tratarse de un corpus común a un área, como es la descripción estructurada de las asignaturas de una universidad, tenderán a repetirse con gran impacto términos característicos del dominio (en nuestro caso, términos estructurales como “tema”, “bloque”, “apartado”, etc.). De hecho, en la tabla 3.8 se muestran los cinco términos que ocurren con mayor frecuencia en la colección de guías objeto de estudio, y sus frecuencias de colección asociadas.

Otro punto interesante a estudiar, es la **distribución de términos entre las guías** de la colección: es decir, cuál es el número de guías entre las que se reparte la cantidad total de apariciones de un término en el corpus.

[22] Aplicando nuevamente la *regla de Sturges* para decidir cuántos intervalos estudiar.

Término	**Frecuencia**
tema	10855
asignatura	2394
bloque	2359
sistema	2270
contenido	2219

Tabla 3.8: Frecuencia de aparición de los 5 términos con mayor presencia en el corpus

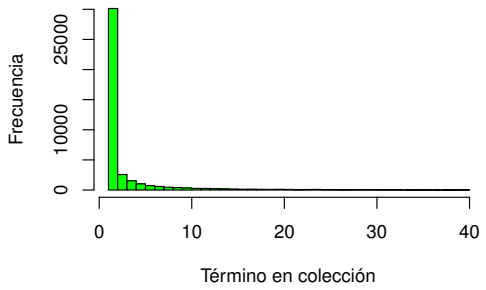


Figura 3.11: Frecuencia de términos en el corpus

Como hemos visto anteriormente, en el corpus de guías hay un total de 445391 términos repartidos en 3143 guías. Un 57.6 % de éstos aparece una única vez en el corpus, por lo que es evidente, que cada uno de ellos aparecerá como máximo en una guía. Pero, ¿cómo se reparten el resto de los términos?

Independientemente del número de apariciones que presenten en el corpus, tenemos que un 65.5 % de los términos aparecen repartidos como máximo en una guía. Es decir, un 7.9 % de los términos se repite más de una vez en el corpus y aún así, sólo está presente en una guía (cada uno de ellos en su guía correspondiente). De nuevo, observando los cuantiles en la tabla 3.9 vemos que existe una concentración de valores extremos en C_{96}.

Ampliando un poco más los datos sobre esta distribución, se obtiene una frecuencia media de aparición de 6.02 guías por término, una mediana de 1 guía por término y una desviación típica de 29.84 guías. Además, también se puede recurrir al histograma de la figura 3.12 para explorar estos datos.

Al igual que sucedía anteriormente, los términos que en más guías están presentes son términos específicos del dominio de estudio. En la tabla 3.10 se muestran los 5 términos que en mayor cantidad de guías se distribuyen.

Dada la gran cantidad de términos que aparecen en una sóla guía, sumado a que sólo un 20 % de éstos aparecen en más de 3 guías, y que un 4 % de los términos se reparte entre 25 y 1500 de estos documentos, se puede pensar que existe gran cantidad de términos específicos de cada guía,

0 %	**4 %**	**8 %**	**12 %**	**16 %**	**20 %**	**24 %**	**28 %**	**32 %**	**36 %**	**40 %**	**44 %**	**48 %**
1	1	1	1	1	1	1	1	1	1	1	1	1
52 %	**56 %**	**60 %**	**64 %**	**68 %**	**72 %**	**76 %**	**80 %**	**84 %**	**88 %**	**92 %**	**96 %**	**100 %**
1	1	1	1	2	2	2	3	4	6	10	25	1523

Tabla 3.9: Cuantiles para la distribución terminológica entre las guías del corpus

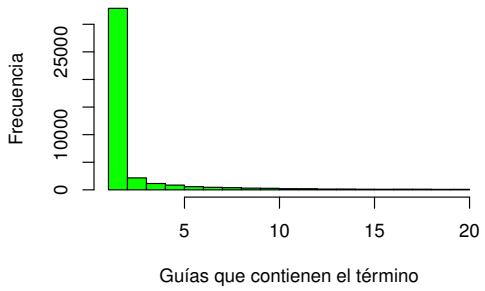


Figura 3.12: Distribución terminológica entre las guías del corpus

Término	**Guías**
asignatura	1523
tema	1414
contenido	1300
estudio	967
concepto	837

Tabla 3.10: Términos con mayor distribución entre las guías del corpus

y una cantidad creciente de éstos que pueden estar representando, desde el área de conocimiento donde se enmarca la guía (para aquellos términos que aparecen en una considerable, pero no demasiado alta, cantidad de guías), hasta la estructura en las que se definen las guías (aquellos términos, como "asignatura" o "tema", que se repiten en una enorme cantidad de guías).

Es lógico pensar que a medida que crece la frecuencia con la que aparece un término en el corpus, aparecerá en mayor cantidad de guías (crecerá su distribución entre éstas). Dicha característica se pasa a comprobar a continuación.

Tras varios test de normalidad sobre las variables "frecuencia de un término en el corpus" y "distribución de un término entre las guías del corpus", podemos rechazar que éstas sigan una distribución normal. Si aplicamos el coeficiente de correlación de *Spearman* obtenemos una asociación positiva al 90 %. Es intuitivo pensar que los datos en la correlación pueden venir condicionados, porque un 57.6 % de los datos aparece una única vez en el corpus y por tanto en una sola guía. Truncando aquellos datos que aparecen una sola vez en el corpus obtenemos una asociación positiva al 85 %. Además, si truncamos todos los valores excepto los más altos (tres últimos cuantiles de la tabla 3.7) la asociación positiva entre las variables se mantiene en valores altos, en torno al 90 %.

No obstante, para completar el análisis, sería conveniente conocer cuáles son las frecuencias medias y máximas de aparición de los términos en cada una de las guías, y si estos valores son independientes, o no, a las longitudes de éstas.

La **frecuencia máxima de aparición de término por documento** viene representada por la frecuencia del término que más se repite dentro cada guía. Dichas frecuencias se distribuyen en el corpus con una media de 9.7 repeticiones por guía, una mediana de 7 repeticiones y una desviación típica de 10. Observando los cuantiles en la tabla 3.11 y apoyándonos en la figura 3.13 vemos que en prácticamente todo el conjunto, las frecuencias máximas varían entre 1 y 25 repeticiones por guía. Sin embargo, en el último cuantil se observa que un 5 % de las guías del corpus tienen frecuencias máximas que van de los 25 a los 231 términos. Parece evidente que este hecho sea consecuencia de una relación entre las variables "frecuencia máxima de término por

0 %	**5 %**	**10 %**	**15 %**	**20 %**	**25 %**	**30 %**
1	2	3	3	4	4	5
35 %	**40 %**	**45 %**	**50 %**	**55 %**	**60 %**	**65 %**
5	6	6	7	8	9	10
70 %	**75 %**	**80 %**	**85 %**	**90 %**	**95 %**	**100 %**
11	12	14	16	19	25	231

Tabla 3.11: Cuantiles para la distribución de máximas frecuencias por guía

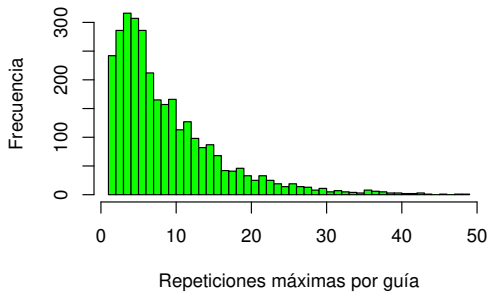


Figura 3.13: Distribución de máximas frecuencias terminológicas por guía

guía" y "longitud en términos de la guía". Más adelante, en este mismo apartado, se estudiará la relación entre dichas variables junto con otra que describa la "frecuencia media de aparición de término por guía".

La **frecuencia media de aparición de término por documento** está representada por el valor promedio de las frecuencias de cada uno de los términos en éste. Estos valores, para el conjunto de las guías del corpus, se distribuyen con una media de 1.54 términos por guía, una mediana de 1.48 términos y una desviación típica de 0.37. Esto nos indica que existe una gran cantidad de términos con frecuencia 1 en las guías, aún en aquellas con mayores frecuencias medias donde, como se puede ver tanto en la tabla 3.12 como en la figura 3.14, la frecuencia media asciende como máximo a casi 6, mientras que la frecuencia máxima de guía más alta en el corpus puede llegar a 231 términos. Por lo anterior se puede deducir que, a pesar de que en muchas guías existan altas repeticiones para algunos términos, las frecuencias medias oscilan entre 1 y 6, por lo que existirán grandes cantidades de términos con frecuencias bajas y muy bajas en los documentos.

Nuevamente, se comprueba que las variables que representan tanto las frecuencias terminológicas media y máxima por guía, como la longitud en cuanto a cantidad términos en éstas, no

0 %	**5 %**	**10 %**	**15 %**	**20 %**	**25 %**	**30 %**
1.000000	1.085714	1.153846	1.204545	1.244227	1.285714	1.333333
35 %	**40 %**	**45 %**	**50 %**	**55 %**	**60 %**	**65 %**
1.367231	1.404762	1.441176	1.479583	1.518589	1.565503	1.609628
70 %	**75 %**	**80 %**	**85 %**	**90 %**	**95 %**	**100 %**
1.657414	1.712688	1.787025	1.883344	2.000000	2.207818	5.956522

Tabla 3.12: Cuantiles para la distribución de frecuencias medias por guía

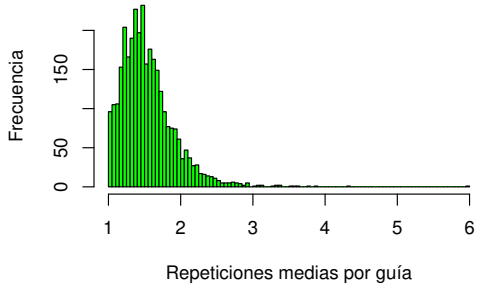


Figura 3.14: Distribución de frecuencias medias de término por guía

	Número de Términos	Frecuencia Máxima de Término	Frecuencia Media de Término
Número de Términos	1.0000000	0.7739876	0.5441779
Frecuencia Máxima de Término	0.7739876	1.0000000	0.7554013
Frecuencia Media de Término	0.5441779	0.7554013	1.0000000

Tabla 3.13: Correlación de *Spearman* entre cantidad y frecuencia de términos por guía

siguen una distribución normal. Tras ello, se aplica el coeficiente de correlación de *Spearman* y obtenemos los valores asociados a la tabla 3.13, donde se puede observar que existe una gran asociación positiva entre frecuencia máxima por guía y longitud de la guía (a mayor longitud de guía, crece la frecuencia máxima de término en ésta). Por otro lado, también se observa una fuerte relación entre las variables que describen la frecuencia media y la frecuencia máxima (parece lógico pensar que las frecuencias medias se vean influidas por las frecuencias máximas para cada documento). Además, también se aprecia una interdependencia positiva, aunque menos fuerte, entre la longitud de las guías y la frecuencia media terminológica en cada una de ellas.

Una vez realizado un análisis exploratorio del corpus de guías (ver apartado 3.2.3), podemos realizar una estimación de cuál será el comportamiento de cada uno de los algoritmos expuestos en el apartado 2.2.

Comportamiento esperado de TF

TF es una técnica local de asignación de pesos, la cual mide la importancia de un término en función de la cantidad de apariciones que tiene en el documento donde reside, es decir, en función de su frecuencia en éste. Dicha función viene definida en la ecuación 2.1.

Como se indicó durante el análisis del corpus, en nuestra colección de guías, los términos se distribuyen con una media de 1.54 términos por guía, con una desviación típica de 0.37. Esto significa que gran cantidad de términos aparecen con frecuencia 1 en las guías.

Anteriormente, comentabamos que gran parte de las guías venían estructuradas por términos comunes, los cuales destacaban tanto por su frecuencia de aparición en el corpus (ver tabla 3.8), como por su distribución en las guías de la colección (ver tabla 3.10). Dichos términos aparecerán con gran frecuencia en gran parte de las guías, por lo que tendrán un gran peso y serán clasificados por TF como términos representativos.

Sin embargo, si obviamos los términos de carácter estructural y los representativos de grandes áreas de conocimiento, TF podría marcar la diferencia entre aquellos términos más importantes y aquellos menos importantes dentro de una guía (sobretodo en las guías más cortas). Aunque esto puede venir bastante condicionado tanto por la longitud, como por el contenido del documento, evidentemente.

Todo parece indicar que TF por sí sólo, no será una buena función para nuestro corpus,

dada la alta aparición de términos estructurales en las guías, así como por la enorme cantidad de términos con frecuencia 1 en el corpus (situación que provoca que la frecuencia de aparición media por guía sea cercana a 1, ver figura:3.14).

Comportamiento esperado de IDF

IDF es una técnica global que mide la importancia de un término en función de la cantidad de documentos en los que aparece en la colección. De esta manera, aquellos términos que se distribuyen en más documentos, obtendrán un peso menor que aquellos que aparecen distribuidos en menor medida, considerando que un término será más importante para un documento si aparece en menos documentos que los demás. La ecuación 2.2 define esta función.

Como nuestro corpus está formado por 3143 documentos, el valor IDF de cada término vendrá expresado en función del número de documentos en los que aparece dicho término. Teniendo en cuenta los datos presentados en la tabla 3.9 y en la figura 3.12, aproximadamente el 65 % de los términos del corpus aparece en una única guía, aproximadamente un 10 % en 2, otro 10 % entre 3 y 10, un 5 % entre 10 y 25, y otro 5 % entre 25 y 1500. Esto quiere decir, que de manera aproximada, un 65 % de los términos compartirá el mismo peso, y sólo un 10 % de los términos del corpus va a presentar un peso que difiera de forma suficientemente significativa de los demás.

Se podría pensar que IDF puede dar, sobretodo, resultados aceptables, en aquellas guías que presenten con gran asiduidad términos con frecuencia 1, sabiendo que la frecuencia media de aparición es de 1.54 términos por guía. Sin embargo, todos los datos vistos nos permiten conocer que dichos términos de frecuencia 1 en documento, también presentan frecuencia 1 en colección por lo que, todo parece apuntar a que IDF por sí sóla no realizará una buena ponderación para la mayor parte de los términos, aunque podría ser bastante útil como herramienta para penalizar términos estructurales.

Comportamiento esperado de TF-IDF

TF-IDF es una de las técnicas de asignación de pesos más conocidas y utilizadas en IR. Combina las funciones TF e IDF vistas con anterioridad, consiguiéndose una mayor efectividad y precisión en los pesos, ya que, de manera local se tiene en consideración la representación de t en el documento (TF), y de manera global se tiene en consideración la representación de t en la colección (IDF). La función de ponderación TF-IDF viene definida en la ecuacion 2.3.

Durante el análisis del corpus de guías (sección 3.2.3), se comprobó que un 65.5 % de los términos aparecían distribuidos únicamente en una guía. También, se constató que un 57.6 % de todos los términos se presentaban con frecuencia 1 en el corpus y, por lo tanto, aparecían en una sóla guía y con frecuencia 1 en ésta. Así que, un 7.9 % del 65.5 % de términos que aparecen únicamente distribuidos en una guía se presenta con frecuencia mayor que 1. Esto quiere decir que con TF-IDF:

1. Un 57.6 % de los términos del corpus presentarán el mismo peso: $log(3143)$
2. Un 7.9 % de los términos son bastante importantes para las guías debido a sus apariciones múltiples en éstas, unido a que, para todo el corpus, aparecen únicamente en su correspondiente documento. Tendrán un peso marcado por su TF, multiplicado por el máximo valor IDF posible de la colección. En esta categoría se encuentran términos que serán, por un lado significativos y por el otro específicos de cada guía, de modo que rescatarlos supondría un logro respecto al objetivo de este trabajo.

Por otro lado, en el corpus existe una fuerte presencia de términos de carácter estructural. Estos términos, a pesar de tener un valor IDF bajo, pueden presentarse fácilmente con frecuencias altas en las guías. Esto puede afectar de manera considerable al peso asignado por TF-IDF, incluso hasta el punto de rescatar entre los términos más destacados, aquellos que por su caracter

Posic.	Término
1	tema
2	transacción
3	sistema
4	dato
5	recuperación
6	aislamiento
7	base
8	consulta
9	procesamiento
10	técnica
11	constar
12	rendimiento
13	almacenamiento
14	arquitectura
15	atomicidad
...	...
172	usar
173	visión
174	vista
175	volcado
176	volátil

Tabla 3.14: TF para 71013041

Posic.	Término
1	aries
2	gestión_de_transacciones
3	granularidad
4	indexacción
5	instantánea
6	introducción_a_la_arquitectura...
7	multiversión
8	recuperación_de_la_información
9	unidad_i._consultas
10	unidad_v
11	unidad_v._desarrollo...
12	volcado
13	atomicidad
14	borrar
15	durabilidad
...	...
172	sistema
173	concepto
174	contenido
175	tema
176	asignatura

Tabla 3.15: IDF para 71013041

Posic.	Término
1	transacción
2	recuperación
3	aislamiento
4	atomicidad
5	consulta
6	procesamiento
7	sistema
8	secuencialidad
9	tema
10	rendimiento
11	dato
12	almacenamiento
13	protocolos
14	bloqueo
15	concurrencia
...	...
172	temático
173	modelo
174	concepto
175	contenido
176	asignatura

Tabla 3.16: TF-IDF para 71013041

0 %	**10 %**	**20 %**	**30 %**	**40 %**	**50 %**	**60 %**	**70 %**	**80 %**	**90 %**	**100 %**
1	1	1	1	1	1	2	2	2	4	26

Tabla 3.17: Distribución de valores TF en la guía 71013041

estructural en las guías son despreciables. Para ilustrar este hecho se tomará una guía cualquiera que contenga el término "tema". En este caso, se ha escogido la asignatura de código 71013041[23], cuyos términos aparecen ordenados según los pesos asignados por las funciones TF, IDF, TF-IDF en las tablas 3.14, 3.15 y 3.16, respectivamente, donde se muestran los 15 términos con los mayores pesos y los 5 con los menores.

El corpus de la guía de código 71013041 está formado por 176 términos, que aparecen en el documento con frecuencias de entre 1 y 8 unidades, excepto 3 términos que aparecen con frecuencias extraordinarias para la guía: "tema", "transacción", y "sistema", con 26, 22, y 16 apariciones, respectivamente. Los términos TF se distribuyen según la tabla 3.17, donde se observa que más de un 50 % de los términos de la guía aparece con frecuencia 1.

En la tabla 3.15, se aprecia que en esta guía aparecen términos estructurales como "contenido", "tema", y "asignatura", que son emplazados en las posiciones más bajas para el *ranking* generado por la función IDF.

Sin embargo, un término como "tema", que aparece en 1414 guías, debería de tener un valor TF-IDF poco significativo, y a pesar de que la función IDF le asigna un peso muy bajo (ver tabla 3.15), aparece en una posición alta en la tabla 3.16 de valores TF-IDF. Esto es consecuencia de que dicho término aparece con una frecuencia muy alta en la guía (26 apariciones), con respecto al resto de términos de ésta (consultar tabla 3.17), presentando una gran varianza respecto a la frecuencia de aparición media de término en la guía de ejemplo (media 2.1 y mediana 1).

TF-IDF puede funcionar bien en muchos casos. Sin embargo, aquellos términos que aparecen con mayor frecuencia y que tienen una fuerte desviación respecto de los valores medios del documento aparecerán con un peso alto, con independencia de que pudieran estar distribuidos en gran cantidad de guías. Por otra parte, tenemos un corpus formado en su 57.6 % por términos que aparecen con frecuencia 1 en la colección, los cuales sólo vendrán determinados por su valor

[23]SISTEMAS DE BASES DE DATOS

Posicición	Término
1	transacción
2	recuperación
3	sistema
4	aislamiento
5	tema
6	atomicidad
7	consulta
8	procesamiento
9	rendimiento
10	dato
11	secuencialidad
12	constar
13	almacenamiento
14	bloqueo
15	protocolos
...	...
172	estudiar
173	modelo
174	concepto
175	contenido
176	asignatura

Tabla 3.18: KLD para 71013041

IDF, independientemente del tamaño y distribución de frecuencias que tengan las guías a las que pertenecen.

Comportamiento esperado de KLD

La función **KLD**, representada mediante la ecuación 2.4, trata de encontrar aquellos términos que aparecen con una probabilidad alta en un documento y, a la vez, con una probabilidad baja en el conjunto total de documentos de la colección. Con ello se pretende extraer aquellos términos que son muy representativos de una guía y no del conjunto total. Para conseguirlo, se basa en las divergencias dadas entre las distribuciones de los términos una determinada guía y las distribuciones de los términos contenidos en el corpus de documentos.

Aplicando esta función parece lógico que todos aquellos términos que son comunes al conjunto de las guías ("tema", "asignatura", etc.) no obtendrán un peso alto, mientras que aquellos términos que sólo aparezcan en una asignatura, se clasificarán como representativos de ésta. Sin embargo, al igual que ocurría con TF-IDF, aquellos términos que difieran considerablemente en frecuencia respecto a los demás de su guía, serán premiados, obteniendo grandes valores KLD. Esto ocurre porque, a pesar de su destacada aparición en el corpus, en términos relativos, la probabilidad de aparición en el documento sigue siendo mucho mayor que la de aparición en el corpus. Esto quiere decir que, con gran probabilidad, KLD extraerá términos estructurales, del mismo modo que lo hacía TF-IDF.

Tomamos de nuevo como ejemplo la guía 71013041 y ordenamos los términos en función del peso asignado por KLD. En la tabla 3.18 se muestran los términos con los 15 mayores pesos y los 5 menores asignados por la función. Se puede observar que, al igual que sucedía con TF-IDF, KLD puede extraer términos estructurales y comunes en la colección como importantes para el documento (p.ej. "tema" aparece como el quinto término más importante de la guía).

Posibles soluciones al problema de la extracción de términos estructurales

Hemos visto que tanto TF-IDF, como KLD pueden extraer términos estructurales de la guía. A continuación, se proponen algunas soluciones para filtrar dichos términos, a la vez que

se pretenden extraer aquellos que sean los más relevantes para cada guía.

- **Solución propuesta a la función TF-IDF:**

  Tal y como se observa en la tabla 3.17, algunas asignaturas presentan términos con altísimas frecuencias respecto al resto de términos de su guía. Esto provoca que la función TF-IDF quede sesgada añadiendo estos valores entre los más importantes, incluso para los valores IDF más bajos. De las diferencias entre las frecuencias medias y máximas por guía se puede inferir que este hecho se repite para gran parte de las guías del corpus (ver tablas 3.11 y 3.12).

  Si atenuamos las frecuencias, de manera que aquellos términos que más apariciones tienen, y que presentan gran desviación respecto al número de apariciones del resto, no sesguen la extracción terminológica con TF-IDF, se podría corregir el problema, adaptando la función a las características de nuestro corpus. Para ello, se propone utilizar la función normalizada **logTF-IDF** definida en la ecuación 3.1.

$$logTF\text{-}IDF(t,d) = log(1 + f_{t,d}) \cdot log\left(\frac{N}{df(t)}\right) \tag{3.1}$$

  Donde $f_{t,d}$ es la frecuencia para el término t en el documento d, N viene dado por la cantidad de documentos del corpus, y $df(t)$ por el número de documentos de éste en los que aparece el término t.

  La utilización de logaritmos presenta una manera de desenfatizar el efecto de la frecuencia [34]. Además, $log(1 + f_{t,d})$ también se puede utilizar como función de pesado local, disminuyendo los efectos provocados por grandes diferencias entre las frecuencias terminológicas [50].

  En la tabla 3.19 se muestran las nuevas extracciones realizadas con esta función para la guía 71013041. Con esta modificación, se espera lograr una mejora en la clasificación de términos estructurales para aquellas guías de características similares a las de la guía ejemplo.

- **Soluciones propuestas a la función KLD:** Al igual que TF-IDF, KLD presenta sesgo hacia aquellos términos que aparecen con alta frecuencia en el documento, incluso cuando se presentan con altas frecuencias en el corpus. Para evitar que esta función extraiga términos estructurales como importantes, se proponen dos soluciones:

  1. Rebajar el peso de los términos en función de su distribución en el corpus. De manera adicional al peso generado por la divergencia obtenida entre la guía y la colección, se pretende penalizar aquellos términos, no sólo que tengan gran frecuencia en el corpus (como hace KLD), sino que aparezcan en gran cantidad de guías. Tal y como se ha indicado anteriormente, IDF puede ser una buena herramienta para llevar a cabo dicha penalización sobre los términos estructurales, debido a las características que presenta nuestro corpus de estudio. Para lograr dicho objetivo, se combinarán las funciones KLD e IDF en la ecuación **KLD-IDF** 3.2.

$$KLD\text{-}IDF(t,d) = P(t,d) \cdot log\left(\frac{P(t,d)}{P(t,C)}\right) \cdot log\left(\frac{N}{df(t)}\right) \tag{3.2}$$

     Donde $P(t,d)$ viene dado por la probabilidad obtenida tras la división de la frecuencia de aparición del término t en el documento d, por la suma de frecuencias de cada uno de los términos del documento d. $P(t,C)$ viene determinado por la probabilidad obtenida tras la división de la frecuencia de aparición del término t en el corpus de todos los documentos C, por la suma de frecuencias de cada uno de los términos

Posicición	Término
1	transacción
2	atomicidad
3	aislamiento
4	secuencialidad
5	consulta
6	recuperación
7	durabilidad
8	monitor
9	protocolos
10	bloqueo
11	hipervínculos
12	procesamiento
13	concurrencia
14	interbloqueo
15	rendimiento
...	...
124	tema
...	...
172	temático
173	modelo
174	concepto
175	contenido
176	asignatura

Tabla 3.19: logTF-IDF para 71013041

del corpus C. N viene dado por la cantidad de documentos del corpus. $df(t)$ por el número de documentos del corpus en los que aparece el término t.

En la tabla 3.20 se muestran los resultados obtenidos con KLD-IDF para la guía 71013041. Al igual que con logTF-IDF, se espera una mejora en la clasificación de términos estructurales.

2. Dar un nuevo enfoque a la función de extracción KLD. En [15] se utiliza la ecuación 2.4, para hallar las divergencias entre la distribución de probabilidad de los términos en la guía y la distribución de probabilidad de los términos en el corpus. Esta última distribución viene determinada por la división del valor de la frecuencia de cada término t en el corpus, entre la suma de frecuencias de todos los términos del corpus.

   Sin embargo, debido a la gran cantidad de términos del corpus (455391 términos), el que aparece con mayor frecuencia en éste ("tema" con 10855 apariciones), aparece "poco" en términos relativos respecto a la frecuencia de aparición que puede tener en una guía. Es decir, que la probabilidad de aparición en la guía puede ser mucho mayor que la del corpus y su KLD presentará un peso alto, a pesar de tratarse del término con mayor número de apariciones de toda la colección.

   Para resolver este problema, se sugiere una nueva aproximación, donde la distribución de referencia no venga dada por la probabilidad de aparición en el corpus como tal, sino como la probabilidad de que el término aparezca distribuido en las diferentes guías del corpus (de manera análoga al objetivo de la función IDF). De esta forma, el término "tema" que aparece en 1414 guías de 3143, obtiene una probabilidad muy alta en relación a la probabilidad de aparición en cualquier guía (con el enfoque anterior la probabilidad de este término en el corpus era del 2.4 % mientras que con este enfoque es del 45 %).

   En este nuevo enfoque se pretende representar la divergencia entre las distribuciones terminológicas de las guías y del corpus, pero desde otra perspectiva. Para ello, se

Posicición	Término
1	transacción
2	atomicidad
3	aislamiento
4	recuperación
5	secuencialidad
6	consulta
7	monitor
8	durabilidad
9	procesamiento
10	protocolos
11	bloqueo
12	hipervínculos
13	rendimiento
14	concurrencia
15	almacenamiento
...	...
80	tema
...	...
172	estudiar
173	concepto
174	modelo
175	asignatura
176	contenido

Tabla 3.20: KLD-IDF para 71013041

propone la ecuación 3.3.

$$KLD^* = P(t,d) \cdot log\left(\frac{P(t,d)}{P(df(t),N)}\right) \tag{3.3}$$

Donde $P(t,d)$ viene dado por la probabilidad obtenida tras la división de la frecuencia de aparición del término t en el documento d, por la suma de frecuencias de cada uno de los términos del documento d. $P(df(t),N)$ viene determinado por la probabilidad obtenida tras la división del número de documentos contenidos en el corpus C en los que aparece el término t, por la cantidad total de documentos en C.

A pesar de que existe una correlación positiva entre la frecuencia de apariciones de un término en el corpus, y su distribución entre las distintas guías, las probabilidades relativas que obtienen los términos estructurales respecto del resto, aumentan considerablemente con este nuevo enfoque y por consiguiente, sus pesos disminuirán. Por otro lado, el resto de extracciones no se deberían de ver afectadas de una manera notable.

En la tabla 3.21 se muestran los resultados con KLD* para la guía 71013041. En dicha tabla vemos que los términos estructurales acaban en las posiciones más bajas de la tabla, independientemente de si se presentan con una fuerte presencia en el documento.

- **Otras soluciones propuestas:** tras el análisis detallado del corpus de guías nos hemos encontrado con una colección en la que las guías suelen ser cortas y contener términos que aparecen con frecuencias medias de entre 1 y 2 repeticiones por guía. Incluso la gran mayoría de los términos aparecen de manera única en el corpus, siendo pocos los que tienen mayores frecuencias o aparecen más distribuidos a lo largo de todas las guías. Además, en nuestro corpus suelen aparecer términos como "tema" y otros característicos de la descripción de la estructura de las guías, que parecen repetirse con gran frecuencia

Posición	Término
1	transacción
2	atomicidad
3	secuencialidad
4	aislamiento
5	durabilidad
6	monitor
7	hipervínculos
8	interbloqueo
9	protocolos
10	bloqueo
11	aries
12	gestión_de_transacciones
13	granularidad
14	indexacción
15	instantánea
...	...
172	dato
173	base
174	técnica
175	sistema
176	tema

Tabla 3.21: KLD* para 71013041

(hemos visto que los términos que aparecen entre los más repetidos en la colección suelen ser términos estructurales, ver tabla 3.8, y que éstos son también los que en más guías aparecen, ver tabla 3.10).

Dadas las características del corpus, en este mismo apartado se ha sugerido el uso de algunas funciones que tienen la finalidad última de penalizar términos característicos de la estructura de la guía. Sin embargo, teniendo en cuenta que los términos que queremos descartar son términos que se repiten con gran asiduidad en el corpus, podemos pensar como solución alternativa, en descartar aquellos que aparezcan en mayor cantidad de guías. De este modo se realizaría una limpieza del corpus en lugar de llevar a cabo un ajuste de las funciones clásicas, con el fin de descartar aquellos términos característicos de la descripción de las guías (a mayor número de guías donde aparece un término, menos importante será para la guía). Con esto se pretende llevar a cabo un experimento complementario que nos permita concluir si realizando una limpieza terminológica y posterior pesado con las funciones clásicas, se obtendrían mejores resultados que evaluando el corpus completo de guías con cualquiera de las funciones propuestas, clásicas o ajustadas.

Para establecer un umbral de limpieza recurrimos, por un lado a la tabla 3.9, donde se observa que para un 4 % de los términos se produce una variación de apariciones por término de entre 25 y 1523 guías, y por otro lado a la tabla 3.7, donde otro 4 % de los términos aparecen en el corpus con frecuencias que oscilan entre 42 y 10855 repeticiones. Es lógico pensar que en la intersección entre los términos que en mayor número de guías aparecen y los términos más frecuentes del corpus se encuentren, tanto aquellos que son propios de la estructura de las guías (como “tema” o “asignatura”), como aquellos que son comunes a un determinado área de conocimiento (como “derecho” o “social”), ya que ambos tipos de terminología aparecerán con total seguridad, en gran cantidad de guías.

Con el fin de limpiar el corpus de términos propios de la estructura de la guía o aquellos comunes a grandes áreas de conocimiento, se opta por retirar los términos que aparecen al mismo tiempo entre los más frecuentes del corpus (4 % representado en el último cuantil de la tabla 3.7) y entre los que en mayor cantidad de guías aparecen (4 % representado en el

último cuantil de la tabla 3.9), para posteriormente llevar a cabo un pesado terminológico sobre el corpus con las funciones clásicas presentadas en este trabajo (TF, IDF, TF-IDF y KLD).

3.2.4. Poblado del *datastore* inicial con los términos extraídos

Tras la ponderación terminológica de cada guía del corpus se procede al poblado automático del *datastore* con los términos que han obtenido mayores pesos. Previamente, se ha creado un fichero (lemas.csv) donde cada línea presenta los términos extraídos para cada asignatura ordenados de mayor a menor según su peso: *código de asignatura ; keyword1 ; keyword2 ; keyword3 ; ...*

De nuevo, para el poblado con los términos contenidos en el citado fichero, hemos utilizado Jena (ver código 3.3, donde se debe establecer de antemano el número de términos a extraer asignándole un valor a la variable "numTerminos"), no sin antes crear la nueva propiedad "*keywords*" que tiene como dominio de partida la clase "Asignatura" y como rango el tipo de datos "cadena de caracteres" (xsd:string).

Código 3.3: Enriquecimiento del *datastore* con la termínologia extraída

```
89  //Creando la nueva propiedad keyword
90  DatatypeProperty keywordP = model.createDatatypeProperty(ontuned + "
        keyword");
91  keywordP.addDomain(asignaturaR);
92  keywordP.addRange(XSD.xstring);
93
94  /*Poblando del nuevo campo keyword con los terminos extraidos
95  y que estan contenidos en el fichero "lemas.csv"*/
96  InputStreamReader inputStreamReader = new InputStreamReader(new
        FileInputStream("lemas.csv"), "UTF8");
97  reader = new CsvReader(inputStreamReader);
98  reader.setDelimiter(';');
99  reader.readHeaders();
100 while (reader.readRecord()) {
101     for (Iterator it = asignaturaR.listInstances(true); it.hasNext();) {
102         Individual ind = (Individual) it.next();
103         if (ind.isIndividual()) {
104             if (ind.getPropertyValue(codigoAsigP) != null) {
105                 String cod = ind.getPropertyValue(codigoAsigP).toString()
                        ;
106                 if (cod.equals(reader.get(0))) {//comparo el codigo
107                     for (int j = 1; j < numTerminos; j++) {
108                         String palIt = reader.get(j);
109                         if ((!palIt.equals("")) && palIt != null) {
110                             //le asigno los keywords a la asignatura
111                             ind.addProperty(keywordP, model.
                                    createTypedLiteral(palIt));
112                         }
113                     }
114                 }
115             }
116         }
117     }
118 }
```

3.3. Discusión integrada de ontologías relevantes para describir procesos universitarios

En este trabajo se ha generado un *datastore* de partida con un vocabulario propio y plano, para después poblarlo con nuevos datos extraídos desde texto libre con técnicas de Minería de Textos. Con el fin de enriquecer el producto que hemos obtenido, se plantean dos líneas:

1. Ampliación del vocabulario mediante etiquetas con el fin de alojar nuevo conocimiento (tal y como se ha hecho con las *keywords*).

2. Disponer el *datastore* para su explotación por parte de terceros.

En este apartado, el cual se concibe como una descripción ampliada de trabajos futuros, vamos a centrarnos en la segunda línea, para lo cual, quizá lo más operativo sea redefinir el vocabulario en términos de ontologías externas más reconocidas, utilizando sus clases, relaciones y propiedades de manera directa o especializada[24].

Aquí se presentan dos niveles de profundidad. Un primer nivel donde la reorientación de nuestros vocabularios se ceñiría justamente a las ontologías mínimas que necesitemos para la representación de los conceptos que hemos abordado en este trabajo. Y un segundo nivel donde se intente llevar a cabo una categorización más profunda de los procesos que ocurren en la universidad, especializando desde arriba y llegando a la selección del etiquetado con mayor conocimiento de causa y de una manera que fomente su reutilización.

Observando los procesos universitarios en general (de investigación, de docencia, de gestión interna, etc.) vemos que son realizados por personas u organizaciones (agentes), que además pueden llevar asociada una descripción, y que existe una serie de entidades diversas que, o bien los consumen o bien los generan (p.ej. aplicaciones *software*).

Desde esta perspectiva, el siguiente objetivo se centra en encontrar aquellas ontologías externas que mejor describan el marco universitario. Para ello, podemos recurrir a la realización de consultas al repositorio de vocabularios puesto en explotación desde Linked Open Vocabularies[25] o también al sitio web de Linked Universities (alianza de la que ya se habló en el apartado 2.1.2 de esta memoria).

Tras revisar los vocabularios, nos hemos encontrado con distintas ontologías dirigidas a distintos objetivos: unas, describen oportunidades de aprendizaje para los estudiantes; otras, procesos docentes; otras, estructuras académicas; etc. (ver apartado 2.1.2). Toda esta fragmentación provoca que existan clases que son redefinidas en los distintos vocabularios, como por ejemplo la clase ":Departmento" (que aparece como "Department" tanto en AIISO como en Bowlogna) y la clase ":Profesor" (que aparece como "Teacher" en "TEACH" y como "Proffesor" en Bowlogna).

Exceptuando los vocabularios VIVO y BIBO (ver apartado 2.1.2), que están bastante consolidados en el terreno de la investigación, existe una dificultad en la puesta en marcha de una integración entre vocabularios propios de un dominio como puede ser el universitario. Esta dificultad viene dada por dos factores:

1. La alta definición de clases como primitivas: un Departamento no debería definirse como clase primitiva, sino como una suborganización que pertenece a otra organización mayor, como puede ser la Facultad y ésta como suborganización de otras, como puede ser la Universidad, etc.

2. La gran cantidad de referencias que parten de las clases, hace difícil que distintas ontologías que tengan conceptos comunes a un dominio puedan ser integradas sin crear inconsistencias o incumplir restricciones.

[24] Podemos definir nuestras clases, relaciones y propiedades como subclases, subrelaciones y subpropiedades de otras existentes para vocabularios externos reconocidos.

[25] `lov.okfn.org`

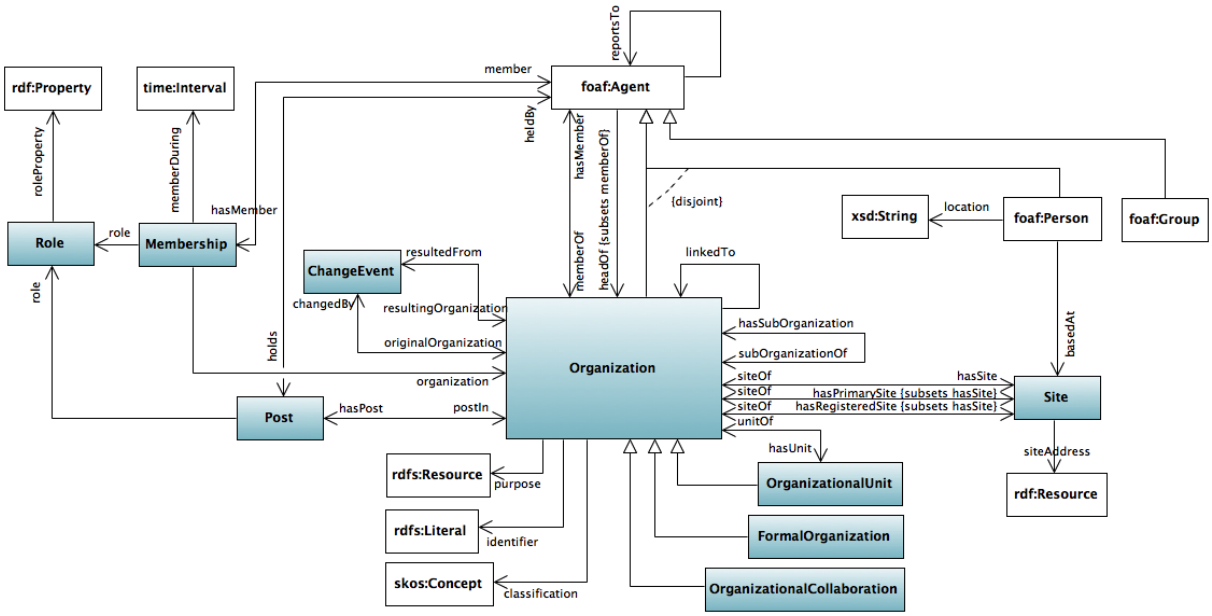


Figura 3.15: Esquema de las relaciones existentes en la ontología ORG

Con la dificultad inherente a la integración de estos vocabularios, surge la necesidad de establecer un marco conceptual más integrador que no es objeto de esta memoria, pero sobre el que se pretende reflexionar a lo largo de este apartado.

ORG es un vocabulario que se presenta como recomendación del W3C con el fin de describir estructuras organizativas a lo largo de gran variedad de dominios (desde, por ejemplo, instituciones gubernamentales, hasta pequeñas empresas) [17]. Además, ORG está conectado a otro vocabulario de gran relevancia, capaz de describir características de personas y grupos sociales: FOAF [13].

Definir la clases “:Universidad”, “:CentroAcadémico” o “:Departamento” como subclases de “org:Organization” permite heredar toda la riqueza de especializaciones y propiedades de ORG; por ejemplo, definir “:Universidad” como “org:FormalOrganization” y “:CentroAcadémico” y “:Departamento” como sus “org:OrganizationalUnit”. Además, permite especializar relaciones como “org:hasUnit” para producir relaciones propias “:tieneCentro” o “:tieneDepartamento”, ambas con dominio “:Universidad” y rangos “:CentroAcadémico” y “:Departamento” respectivamente. De forma análoga se pueden definir sus respectivos puestos de dirección (rector, decano, director...) genéricamente, a falta de instanciación personal, a través especializaciones de la relación “org:hasPost”.

La asignación de roles y pertenencia en ORG es muy versátil, a través de relaciones n-arias que permiten asignar el rol asumido por un agente en un periodo determinado. Y estos agentes pueden ser organizaciones (“org:Organization”), grupos de personas (“foaf:Group”) o personas (“foaf:Person”). Una mera revisión del gráfico 3.15, que describe ORG, permite apreciar otras facilidades ontológicas que pueden heredarse.

Esta integración entre ORG y FOAF invita a modelar el personal académico como especialización de FOAF. En particular, los equipos docentes o los grupos de estudiantes que tienen asignada un tarea común se pueden modelar a través de “foaf:Group”.

Esta perspectiva, brevemente enunciada, permite modelar de forma integrada todos los agentes implicados en procesos universitarios. A nivel organizativo, porque las estructuras académicas se pueden complementar con las unidades de soporte técnico y administrativo, con la misma ontología superior de referencia. Y a nivel de personas o grupos, de la misma manera, porque se integran profesores, estudiantes y personal de soporte (junto a los grupos conjuntos que se quieran definir).

Por otro lado, en la literatura existe un vacío en la descripción ontológica de procesos académicos en general, a excepción de los procesos de investigación [28], donde mediante el vocabulario PROV-O [37] se describe el recorrido de los datos durante el proceso. Además, con el modelado de procesos surgió la necesidad de desarrollar una ampliación del vocabulario de manera que, no sólo permitiera la descripción de procesos en pasado, sino que permitiera la descripción de un plan a seguir para llevar a cabo su ejecución a futuro: P-Plan [27].

Dado este vacío, sumado al amplio recorrido del modelado de procesos ya existente en el área de investigación, podemos pensar en modelar los diversos procesos académicos que se puedan dar en la universidad.

Ninguno de las vocabularios similares existentes (en lov.okfn.org) permite una descripción detallada de procesos. No al menos con el detalle que permite PROV-O. Quizá porque esas ontologías son previas a esta recomendación del W3C.

Sí que hay bastantes ontologías aplicables a la descripción de los recursos (las entidades) utilizadas o producidas: guías didácticas, vídeos, audio, artículos de investigación, tareas propuestas, tareas entregadas... Estos vocabularios permiten la descripción de entidades respecto a varias perspectivas, entre otras posibles: su formato, la traza de su autoría, el papel que juegan las entidades respecto al tipo de proceso, su relación con taxonomías o tesauros que las agrupan temáticamente...

Para formato y trazas de autoría se utiliza de forma generalizada Dublin Core [31]. El papel de estas entidades en los diversos procesos académicos se puede presentar de forma más integrada si estos procesos se modelan como especializaciones de PROV-O o de su ampliación P-Plan.

Por último, la categorización terminológica y temática de las entidades se está abordando mediante vocabularios SKOS especializados [43]. Estos vocabularios amplian las relaciones entre clases y propiedades que facilita RDFS [12], sin llegar a la rigidez de las ontologías OWL [22], y, sobre todo, permite un amplio conjunto de anotaciones, descripciones y comentarios sobre los esas clases y propiedades.

En esta memoria, esta perspectiva temática de las entidades es particularmente relevante. El enriquecimiento del *datastore*, por importación de información desde fuentes no estructuradas, puede facilitar la clasificación temática de recursos referidos en el datastore o puede asociar a estos recursos toda una colección de términos estructurados respecto a tesauros o taxonomías que pueden emergen de estos procesos de enriquecimiento.

Capítulo 4

Evaluación del Experimento y Análisis de Resultados

4.1. Resultados de la construcción del *datastore* a partir de fuentes estructuradas

Como primer resultado del trabajo hemos obtenido un *datastore* con 100617 tripletas, que puede ser consultado, modificado o ampliado por agentes externos.

Con el fin de establecer un punto de consulta a nuestro *datastore* se puede hacer uso del servidor Apache Jena-Fuseki [1], el cual proporciona los servicios de consulta y actualización de tripletas RDF mediante SPARQL sobre protocolo HTTP.

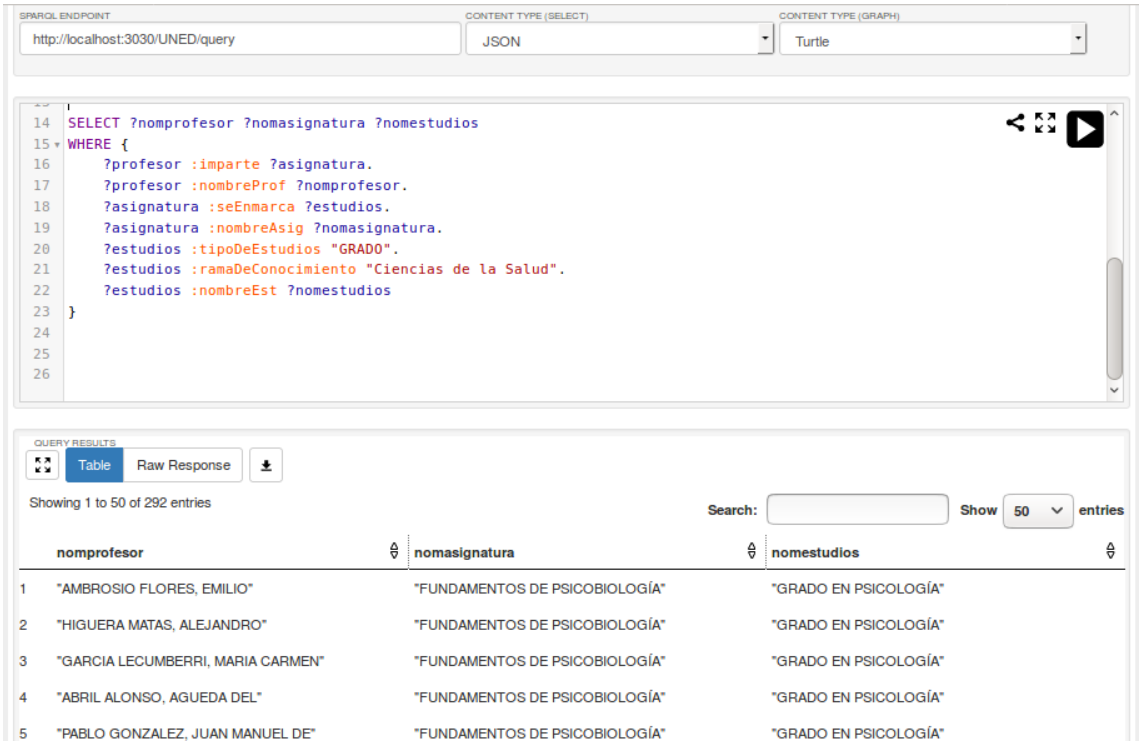


Figura 4.1: Consulta a través de la interfaz web de Fuseki

En la figura 4.1 se puede observar una consulta realizada a través de la interfaz web que ofrece el servidor Fuseki, donde se obtiene un listado con los nombres de los profesores, las

[1] `https://jena.apache.org/documentation/serving_data/`

asignaturas que tienen asignadas y el plan de estudios al que pertenecen dichas asignaturas, de todos los Grados de la rama de conocimiento de "Ciencias de la Salud".

A continuación, se presentan dos ejemplos de consulta hacia el *datastore* generado inicialmente. Dichas consultas pueden realizarse en el SPARQL *endpoint* ofrecido por Fuseki:

1. Se desea conocer el nombre de la asignatura de código "61044135". Con el propósito de obtener esta información se podría realizar una consulta SPARQL como la que se muestra en el código 4.1, obteniendo como resultados los mostrados en la tabla 4.1.

Código 4.1: Consulta 1 al SPARQL Endpoint

```
119
120 PREFIX  :      <http://localhost/ontouned/ontologia/uned#>
121
122 SELECT distinct?NombreAsignatura
123 WHERE {
124         ?asignatura :nombreAsig ?NombreAsignatura.
125         ?asignatura :codigoAsig "61044135".
126 }
```

NombreAsignatura
RELATIVIDAD GENERAL

Tabla 4.1: Resultados a la consulta 1

2. Se desea consultar el nombre de los profesores que imparten la asignatura "BASES DE LA INGENIERÍA AMBIENTAL" y su departamento y facultad asociados. Para ello se puede realizar una consulta SPARQL como la presentada en el código 4.2, obteniendo como resultados los presentados en la tabla 4.2.

Código 4.2: Consulta 2 al SPARQL Endpoint

```
129
130 PREFIX  :      <http://localhost/ontouned/ontologia/uned#>
131
132 SELECT distinct?NombreFacultad ?NombreDepartamento ?NombreProfesor
133 WHERE {
134         ?asignatura :nombreAsig "BASES DE LA INGENIERIA AMBIENTAL".
135         ?asignatura :esImpartida ?profesor.
136     ?profesor :nombreProf ?NombreProfesor.
137         ?profesor :esPersonalDe ?departamento.
138         ?departamento :nombreDep ?NombreDepartamento.
139     ?departamento :formaParteDe ?facultad.
140     ?facultad :nombreFacul ?NombreFacultad.
141 }
```

NombreFacultad	NombreDepartamento	NombreProfesor
FACULTAD DE CIENCIAS	QUÍMICA INORGÁNICA Y QUÍMICA TÉCNICA	ALVAREZ RODRIGUEZ, JESUS
FACULTAD DE CIENCIAS	QUÍMICA INORGÁNICA Y QUÍMICA TÉCNICA	MAROTO VALIENTE, ANGEL
FACULTAD DE CIENCIAS	QUÍMICA INORGÁNICA Y QUÍMICA TÉCNICA	MUÑOZ ANDRES, VICENTA

Tabla 4.2: Resultados a la consulta 2

4.2. Resultados de la extracción terminológica y del enriquecimiento del *datastore*

Este apartado, que tiene como objetivo la revisión de resultados obtenidos por la extracción terminológica de las guías y su utilización para el enriquecimiento en el poblado de la ontología, se divide en una serie de subapartados secuenciales:

1. Preparación del proceso de evaluación

   a) Métodos utilizados para la evaluación: breve descripción de las métricas que se utilizarán para llevar a cabo la evaluación de la extracción terminológica.

   b) Plataforma para la obtención del fichero *Gold Standard*: presentación de la interfaz a través de la cual se genera el *Gold Standard* que se utilizará para realizar la evaluación de las distintas modalidades de extracción terminológica.

   c) Recogida y análisis del *Gold Standard*: descripción del *Gold Standard* recopilado.

2. Resultados de la evaluación: análisis mediante las distintas métricas de evaluación de los resultados obtenidos mediante las diferentes funciones y métodos en la asignación de pesos a los términos.

3. Resultados del enriquecimiento del *datastore* con palabras clave: ampliación del potencial de consulta con el *datastore* resultante enriquecido.

4.2.1. Preparación del proceso de evaluación

Métodos utilizados para la evaluación

Hasta ahora, disponemos de un corpus en el que los términos de cada documento tienen asociado un peso que ha sido determinado por alguna función. Dichos pesos establecen un orden de representatividad entre los términos de un documento determinado de manera que, si los ordenamos según su peso en orden descendente, en las primeras posiciones del *ranking* deberían quedar aquellos más representativos de cada documento.

Para la evaluación de estas funciones se tendrá en cuenta la efectividad en cuanto a que, entre las primeras posiciones del *ranking* para cada documento, se encuentren aquellos términos más importantes en relación al contenido de la guía, reduciendo al mínimo el número de términos despreciables.

Con el objetivo de llevar a cabo la evaluación se tomará una colección de guías de referencia. Dicha colección se formará a partir de una cantidad suficiente de guías, las cuales contendrán los términos más importantes proporcionados por los autores de las guías para cada una de ellas. Para llevar a cabo la creación de este corpus de referencia (*Gold Standard*), se solicitará la participación de los profesores que diseñaron las guías de modo que, seleccionen aquellos términos que consideren más importantes de éstas. Dicho proceso será llevado a cabo mediante una interfaz descrita más adelante, en este mismo apartado.

Más adelante describiremos las diferentes métricas utilizadas para llevar a cabo la evaluación, las cuales pueden utilizarse de manera individual sobre una guía determinada, o sobre un conjunto de éstas, mediante un promedio de los valores obtenidos para cada guía del conjunto[2]. En este trabajo, la evaluación se realizará sobre todo el conjunto de guías que forman el *Gold Standard*.

Antes de estudiar las diferentes formas de evaluación, conviene echar un vistazo a la figura 4.2, donde se puede observar la intersección entre dos conjuntos o bolsas de términos. Por un lado, se tiene el conjunto X de términos extraídos para una guía, mediante alguna de las funciones

[2]Excepto para la medida *MAP*

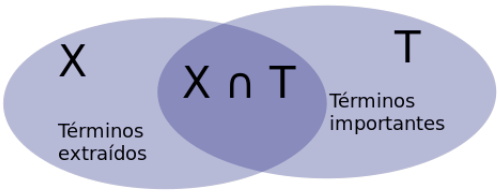


Figura 4.2: Términos extraídos vs términos importantes

Términos	Contenidos en X	No contenidos en X
Contenidos en T	VP	FN
No Contenidos en T	FP	VN

Tabla 4.3: Matriz de confusión para un sistema de extracción de términos

de pesado propuestas. Por otro lado, tenemos el conjunto T de los términos importantes para dicha guía, seleccionados por el personal docente encargado de la elaboración de la guía. En la intersección entre ambos conjuntos $X \cap T$, se encuentran aquellos términos que son importantes y al mismo tiempo extraídos por el sistema automático de extracción.

A partir de la distribución de los términos en la figura 4.2 se obtiene la matriz de confusión representada en la tabla 4.3, cuyos grupos vienen determinados por:

- *VP* o *Verdadero Positivo*: conjunto formado por aquellos términos que aparecen en X y en T, en la figura 4.2 $X \cap T$. Es decir, aquellos que siendo importantes, han sido extraídos mediante el sistema.

- *FP* o *Falso Positivo*: conjunto formado por aquellos términos que ha extraído el sistema, pero que no se encuentran entre los términos más importantes en la guía. Es decir, aquellos términos contenidos en X que no están contenidos en T.

- *VN* o *Verdadero Negativo*: conjunto formado por aquellos términos que contiene la guía y que no aparecen ni en X ni en T.

- *FN* o *Falso Negativo*: conjunto formado por aquellos términos que son importantes y no ha extraído el sistema. Dicho conjunto lo forman los términos que aparecen en T pero no en X.

A continuación se describen algunas de las metodologías más utilizadas como métricas de evaluación en IR [40]:

- **Exactitud:** Representa la fracción de términos importantes de entre todos los extraídos. Teniendo en cuenta la matriz de confusión (tabla 4.3), la exactitud queda definida según la ecuación 4.1.

$$exactitud = \frac{VP + VN}{VP + VN + FP + FN} \tag{4.1}$$

  Normalmente, esta métrica no resulta ser fiable, dado que la gran mayoría de los términos están contenidos fuera de la categoría T de términos importantes. Dicha característica hace que la métrica quede sesgada, dado que normalmente, en documentos extensos, VN va a ser sustancialmente superior a la suma de VP, FP y FN. Es por ello que, la métrica de exactitud se suele utilizar con resultados fidedignos en conjuntos donde el número de términos en VN y VP sean equitativos.

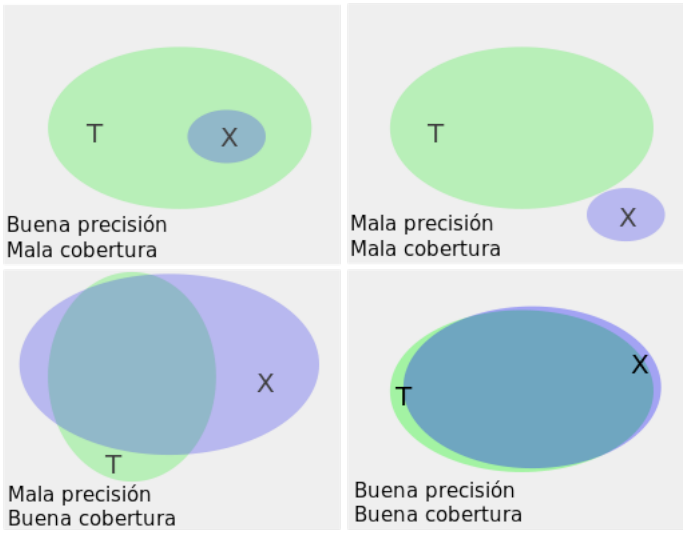


Figura 4.3: Escenarios posibles en precisión y cobertura

- **Precisión y Cobertura:** son las métricas más utilizadas para evaluación en IR. La precisión, representada mediante la ecuación 4.2, mide la capacidad del sistema para extraer exclusivamente términos importantes, mientras que la cobertura, representada mediante la ecuación 4.3, mide la capacidad del mismo para extraer todos los términos que son importantes. En nuestro caso los términos extraídos se corresponden con los documentos recuperados en un sistema de IR. La figura 4.3 muestra los posibles escenarios que pueden darse en un sistema IR en concepto de precisión y cobertura, siendo X el conjunto de términos extraídos y T el conjunto de términos importantes.

$$precisión = \frac{VP}{VP + FP} \tag{4.2}$$

$$cobertura = \frac{VP}{VP + FN} \tag{4.3}$$

  Lo habitual en un sistema de evaluación de este tipo es que precisión y cobertura guarden una relación inversamente proporcional, es decir, a mayor cobertura obtenida menor será la precisión y viceversa. Una métrica muy utilizada en IR, que auna precisión y cobertura en un único valor que estima la bondad del sistema, es la **medida-F** que se define en la ecuación 4.4, donde β es un factor que permite asignar más ponderación a la precisión o a la cobertura según se desee. Para lograr una métrica armónica entre precisión y cobertura, se deberá asignar a β el valor 1. Si lo que se desea es dar mayor importancia a la cobertura, β deberá tener un valor mayor a 1. Y por el contrario, si se desea que la precisión tenga más peso que la cobertura, el valor de β deberá ser inferior a 1.

$$Medida\text{-}F_\beta = (1 + \beta^2) \cdot \frac{precisión \cdot cobertura}{(\beta^2 \cdot precisión) + cobertura} \tag{4.4}$$

- **Precisión a los 11 niveles estándar de cobertura:** consiste en calcular la precisión para los distintos niveles de cobertura, desde la precisión a cobertura 0, escalando en intervalos de 0.1, hasta la precisión a cobertura 1. Dado que puede ocurrir que para un determinado nivel de cobertura no se disponga de su valor exacto de precisión (p. ej. cobertura a 0.2), los niveles de precisión se calculan interpolando los valores a una determinada cobertura, de la manera que especifica la ecuación 4.5. De modo que, la precisión a

un determinado nivel i de cobertura viene determinada por la máxima precisión conocida entre el nivel i-ésimo de cobertura y el nivel (i+1)-ésimo.

$$P(C_i) = max_{C_i \leq C \leq C_{i+1}} P(C) \tag{4.5}$$

- **Precisión a n:** en esta métrica únicamente se obtiene la precisión a los n primeros términos extraídos. Es interesante cuando se busca que el sistema extraiga, entre las primeras posiciones del ranking términos, el mayor número de términos importantes sin necesidad de que se extraigan todos aquellos que son importantes.

- **Precisión media no interpolada:** se calcula como la media de las precisiones obtenidas para cada término importante extraído hasta que se obtengan todos los términos importantes. Esta métrica tiene en cuenta el orden en el que la función de *ranking* ordena los términos, premiando aquellos sistemas que devuelvan los términos importantes en las primeras posiciones. La precisión media no interpolada viene representada en la ecuación 4.6.

$$AveP = \frac{\Sigma_{k=1}^{n}(P(k) \cdot rel(k))}{\text{número de términos importantes}} \tag{4.6}$$

  Donde, $AveP$ es la precisión media no interpolada, k es la posición del término en el *ranking* de términos extraídos por el sistema, $P(k)$ es la precisión obtenida para los k primeros términos extraídos por el sistema, $rel(k)$ es el valor de relevancia del término que aparece en la posición k del *ranking*, con valor 1 si es importante y con valor 0 en otro caso, y "número de términos importantes" viene dado por todos los términos que forman el conjunto T de términos, es decir, por VP y FN.

- **Precisión-R:** en esta métrica de evaluación se obtiene la precisión a los R primeros términos extraídos, siendo R el número de términos importantes.

- **Media de la precisión promedio o MAP:** en esta métrica se hace la media de las precisiones medias no interpoladas para todos los documentos a evaluar. Es una métrica global donde se evalúa la precisión sobre un conjunto de documentos, en lugar de para un documento determinado. La fórmula para la obtención de esta métrica viene representada por la ecuación 4.7.

$$MAP = \frac{\Sigma_{q=1}^{Q}(AveP(q))}{\text{Q}} \tag{4.7}$$

  Donde Q es el conjunto de todos los documentos q a evaluar.

Plataforma para la obtención del fichero *Gold Standard*

Cada año la UNED actualiza las guías de estudio de sus asignaturas, las cuales proporcionan la información necesaria al estudiante, incluyendo orientaciones sobre contenidos y actividades propuestas. Los Equipos Docentes son los encargados de la realización de las Guías de sus asignaturas, por lo que conocen el contenido de éstas con la suficiente profundidad como para hacer una selección adecuada sobre los conceptos que definen de la mejor manera posible dicho contenido.

Para realizar la evaluación del sistema de extracción terminológica se necesita de un fichero *Gold Standard* que contenga aquellos términos importantes contra los que se deberán comparar los términos extraídos por nuestro sistema. Así pues, con el objetivo de generar este *Gold Standard*, se ha solicitado la participación del profesorado para que, mediante una plataforma *online*, los docentes que participen puedan seleccionar los términos más importantes de las páginas de contenidos de sus guías docentes para el curso 2016-2017.

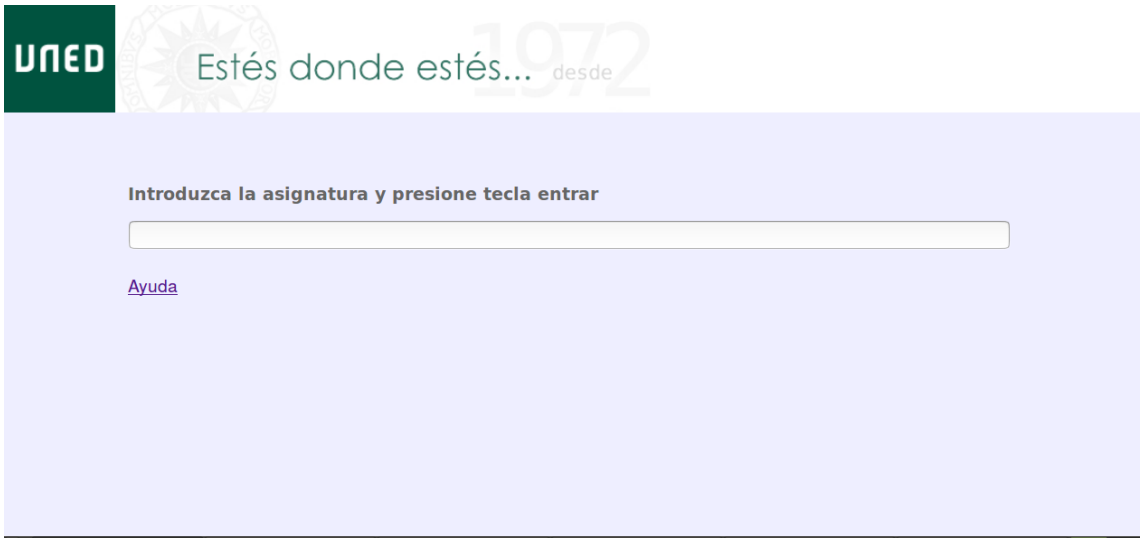


Figura 4.4: Página principal plataforma para la creación del Gold Standard

Figura 4.5: Cuadro de búsqueda por asignatura

La selección de conceptos se llevará a cabo a través de una página web creada a tal efecto. La intención es que el proceso sea lo más sencillo, rápido e intuitivo posible, para conseguir la mayor participación y número de asignaturas posible en el *Gold Standard*, y así crear un fichero de evaluación lo suficientemente grande como para que las pruebas realizadas sean determinantes.

Una vez se encuentre la función de extracción de terminología que más se ajuste al *Gold Standard*, se aplicará al conjunto tortal de asignaturas y podrá entonces realizarse el poblado del capo *keywords* de la clase asignatura de nuestra ontología.

En la figura 4.4 aparece la página principal de la web, la cual dispone únicamente de un enlace de ayuda y de una barra de búsqueda donde al introducir tanto parte del nombre, como del código de la asignatura, se extiende un desplegable con las asignaturas que presentan coincidencias con el criterio de búsqueda (ver figura 4.5).

Una vez seleccionada e introducida la asignatura en el cuadro de búsqueda, el profesor es redirigido a la página específica de su asignatura, donde podrá seleccionar los conceptos clave de ésta (ver figura 4.6). Dichos conceptos han sido extraídos del texto que aparece en la página web de contenidos de la guía de la asignatura en cuestión. Para ello, el texto de cada página de contenidos ha sido automáticamente *tokenizado* y etiquetado morfosintácticamente, extrayéndose los lemas de cada palabra, y eliminándose posteriormente *stopwords* y caracteres especiales.

En la página de selección de términos se muestran 40 conceptos para la asignatura introducida. Estos conceptos han sido previamente ordenados según su frecuencia en la guía para facilitar la localización de aquellos que pudieran ser más importantes, sin sesgar ni condicionar al usuario, como pasaría si se muestran los términos ordenados por alguna función más completa (como KLD o TF-IDF). Este ordenamiento (y no uno aleatorio) se hace para facilitar al docente la localización de términos importantes entre los 40 primeros, sobre todo, en aquellas asignaturas

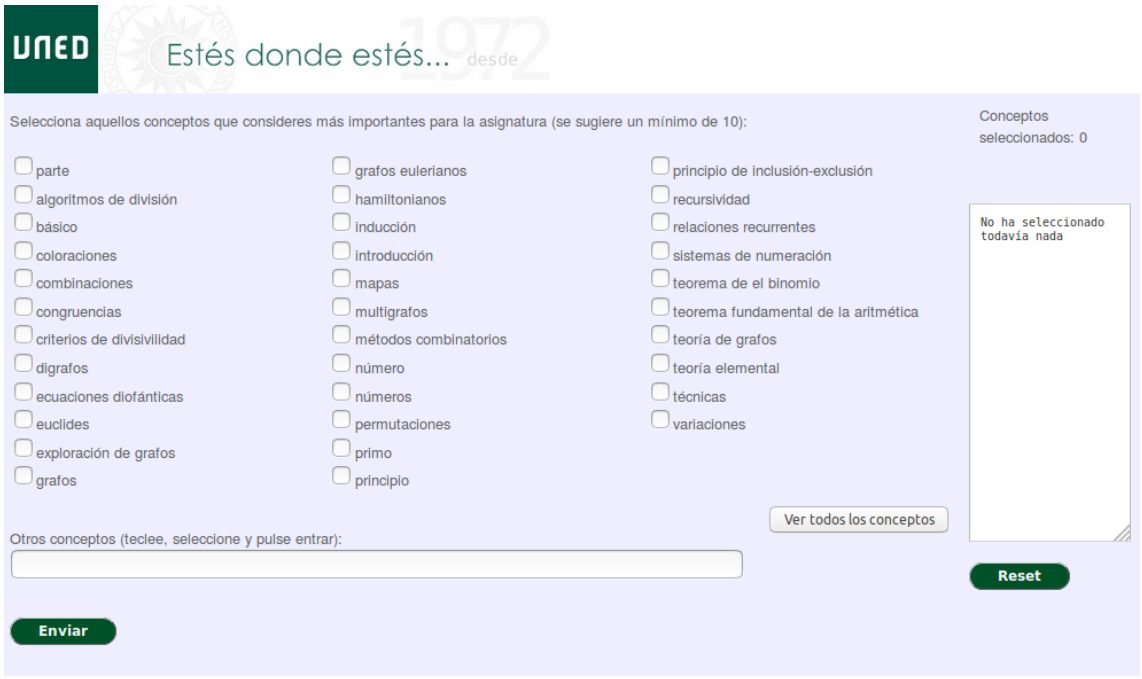


Figura 4.6: Página de selección de conceptos para una guía

con cientos o miles de ellos. No obstante, al ordenarse únicamente por frecuencia en la guía, en las primeras posiciones aparecerán tanto términos importantes como términos no importantes, entre los que será frecuente la localización de términos estructurales. Aquí, hay que recordar que según la bibliografía estudiada en la sección 2.2, para localizar de manera efectiva los términos importantes, es conveniente tener en cuenta su aparición, tanto en la guía, como en el corpus completo de guías (factor local y factor global).

Sin embargo, aunque sólo se muestran los 40 primeros términos de cada guía, el docente puede escoger entre todos los que aparecen contenidos en su asignatura. Para hacer esto posible, se han incluido dos mecanismos: el botón de "*Ver todos los conceptos*" y un cuadro de búsqueda de términos.

El botón de "*Ver todos los conceptos*", tal y como indica su etiqueta, al ser presionado mostrará todos los términos de la asignatura. Pero el hecho de que aparezcan todos los términos de una asignatura que contenga cientos o miles de éstos, puede convertir la localización de conceptos en un pasatiempo para el docente. Por esta razón se ha incluido también el cuadro de búsqueda terminológica, herramienta ideada para aquellas guías más largas donde localizar un término a simple vista puede ser costoso. Bastará con introducir en el cuadro el concepto que se busca o parte de él para que, mediante un desplegable se muestren todos los conceptos que coincidan con los criterios de búsqueda introducidos en éste.

Cada concepto que se seleccione por cualquiera de los métodos (haciendo click sobre el concepto o mediante el cuadro de búsqueda terminológica) se mostrará en el cuadro de la izquierda "*Conceptos seleccionados*" de manera que, el docente tendrá una visión sintetizada de cuántos conceptos ha escogido y cuáles han sido.

La plataforma sugiere a los docentes que seleccionen al menos 10 términos, pudiendo éstos marcar libremente la cantidad que consideren oportuna, incluso menos de 10 si consideran que sólo esos términos son representativos para su asignatura.

Una vez marcados los conceptos que el docente considera más importantes para su guía, pulsará el botón de "enviar". Los resultados quedarán entonces automáticamente almacenados, de manera que puedan ser utilizados posteriormente para evaluar el sistema de extracción automática de terminología.

0 %	**6 %**	**12 %**	**18 %**	**24 %**	**30 %**
3	7	10	10	10	10
36 %	**42 %**	**48 %**	**54 %**	**60 %**	**66 %**
11	12	13	14	14	15
72 %	**78 %**	**84 %**	**90 %**	**96 %**	**100 %**
16	18	20	22.2	33	89

Tabla 4.4: Cuantiles para la distribución de frecuencias en el *Gold Standard*

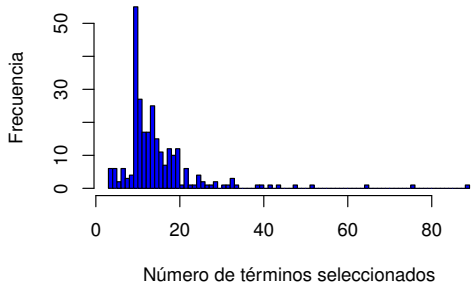


Figura 4.7: Distribución de frecuencias en el *Gold Standard*

De manera adicional, en la página principal aparece un enlace de ayuda que redirige al usuario a una página con información sobre la plataforma, su utilización y sus objetivos.

Recogida y análisis del *Gold Standard*

Tras tres semanas abierta la plataforma a los docentes para la recogida de datos, disponemos de un *Gold Standard* con 314 entradas, entre las cuales se observa que existen ciertas repeticiones para algunas guías. Esto se debe a que los profesores de un mismo Equipo Docente comparten y pueden seleccionar términos de la misma asignatura. En algunas líneas se observa que dichos docentes han seleccionado exactamente los mismos términos, pero en otras existen algunas diferencias. Para evaluar aquellas líneas que hacen referencia a la misma asignatura pero que difieren entre sí, se utilizará la unión entre dichas líneas en una sóla de manera que, en ésta apareceran todos los términos seleccionados por el mismo Equipo Docente (cada término aparecerá una única vez por línea).

Una vez hecho esto, el *Gold Standard* queda listo para poder llevar a cabo las evaluaciones de las distintas funciones de pesado con las métricas descritas, pero antes haremos una pequeña exploración de éste.

Entre las 314 entradas hay un total de 269 guías diferentes, con una media de términos seleccionados por equipo y guía de 15, una mediana de 13 y una desviación típica de 9.76 términos. En la tabla 4.4 se muestra la distribución por cuantiles de estas longitudes, quedando también representada dicha información en la figura 4.7.

Por otra parte, podemos llevar a cabo una exploración por frecuencias de términos seleccionados por los docentes para la elaboración del *Gold Standard*. En nuestro caso, existe una gran población de términos que son escogidos de manera única por los profesores, obteniendo una frecuencia media de 1.69, una mediana de 1, y una desviación típica de 1.79 repeticiones. Además tanto en la figura 4.8 como en la tabla 4.5 se observa que en torno a un 70 % de los términos escogidos por los docentes son seleccionados de manera singular. Sin embargo, en nuestro *Gold*

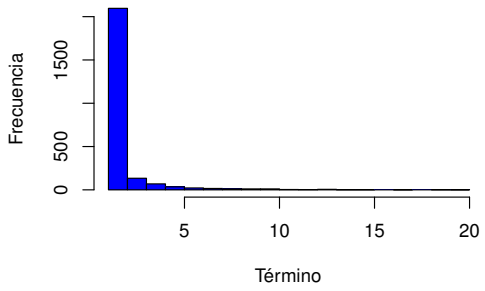


Figura 4.8: Frecuencia de término en el *Gold Standard*

0 %	**5 %**	**10 %**	**15 %**	**20 %**	**25 %**	**30 %**
1	1	1	1	1	1	1
35 %	**40 %**	**45 %**	**50 %**	**55 %**	**60 %**	**65 %**
1	1	1	1	1	1	1
70 %	**75 %**	**80 %**	**85 %**	**90 %**	**95 %**	**100 %**
1	2	2	2	3	5	20

Tabla 4.5: Distribución de frecuencias terminológicas en el *Gold Standard*

Standard también se han rescatado términos que son comunes a varias guías (alguno hasta 20) y que pueden ser importantes no sólo para dichas guías, sino también para el área de conocimiento en la que éstas se enmarcan (se pueden observar algunos de estos términos al inicio de la tabla 4.6).

4.2.2. Resultados de la evaluación

En este apartado se estudiarán los resultados obtenidos tras la evaluación de las distintas metodologías utilizadas. Para un análisis más completo, se mostrarán los resultados ofrecidos por las distintas métricas utilizando el *Gold Standard* como fichero de referencia. En primer lugar, se mostrarán los resultados obtenidos tras la evaluación del corpus completo de términos con las distintas funciones de pesado propuestas (clásicas y ajustadas). En segundo lugar, se estudiarán los resultados obtenidos por las funciones clásicas sobre el corpus libre de aquellos términos que, por su alta frecuencia y distribución, pueden ser considerandos secundarios. Por último, se realizará una comparación entre todos los resultados obtenidos, con el fin de tomar una decisión para realizar el poblado del campo *keywords* de la clase asignatura en nuestro *datastore*. Cabe remarcar que dependiendo del uso que se le quiera dar a las *keywords*, los criteros a seguir pueden ser diferentes a los escogidos en este trabajo (por ejemplo, si en el objetivo primara la cobertura sobre la precisión).

Resultados de la evaluación de las distintas funciones de ponderación

Tras la generación del fichero *Gold Standard* mediante la participación de los profesores en la plataforma descrita en el apartado 4.2.1, se procede a evaluar los términos extraídos de manera automática mediante las funciones presentadas en este trabajo (TF, IDF, TF-IDF, KLD, KLD*, KLD-IDF).

Durante el análisis del corpus en el apartado 3.2.3 estudiamos las particularidades que pre-

Término	**Guías**
sistema	20
investigación	19
modelo	19
estructura	17
historia	15
educación	15
aprendizaje	14
diseño	14
evaluación	13
...	...
democracia representativa	1
federal	1
parlamento	1
república	1
auditoría	1
auditor	1

Tabla 4.6: Términos con mayor distribución en el *Gold Standard*

sentaba la colección de términos y su distribución a lo largo del corpus de documentos. Además en dicho apartado estimábamos el comportamiento que tendrían las distintas funciones sobre este corpus.

En primer lugar presentamos la evaluación de las medidas de precisión, cobertura, y medida-F (armónica) en el progreso de extracción desde uno hasta los veinte primeros términos por parte de las funciones presentadas en este trabajo.

En la figura 4.9 se presenta la gráfica de precisiones donde se ven representadas las líneas que describen la evolución que muestran las distintas funciones en el recorrido mencionado. Así, se puede observar que si sólo se extrajera un término por guía, las mayores precisiones vendrían dadas por la función KLD-IDF y KLD*; sin embargo, la gráfica refleja que a mayor número de términos extraídos, mayor es la perdida de precisión de estas dos funciones en comparación con las clásicas KLD y TF-IDF. Por otro lado, logTF-IDF presenta precisiones ligeramente inferiores a KLD y TF-IDF, pero podría resultar interesante si se deseara descartar los términos estructurales aunque esto supusiera sacrificar levemente la precisión.

KLD*, KLD-IDF, y logTF-IDF fueron propuestas para penalizar términos estructurales que aparecieran con gran frecuencia en las guías (ver apartado 3.2.3). Sin embargo, todo parece indicar que además de penalizar los términos estructurales, también se penalizan otros términos que pueden haber sido considerados como importantes en la guía, hecho que también se ve reflejado si observamos la gráfica de coberturas en la figura 4.10.

No obstante, la bondad de las funciones no puede ser medida únicamente con la métrica de precisión o la de cobertura. Dichas métricas guardan de manera general una relación inversamente proporcional y lo ideal es utilizar una que unifique ambos valores, representando de una manera más confiable la bondad de la función. Para ello, se utiliza la medida-F armónica, que para los 20 primeros términos extraídos viene representado en la gráfica 4.11. En esta figura se puede observar que la función que mejores extracciones obtiene es TF-IDF, con resultados casi idénticos para KLD, ambas funciones seguidas muy de cerca por logTF-IDF y KLD-IDF. TF y KLD* obtienen peores extracciones que el resto de funciones, exceptuando IDF que presenta los peores resultados del experimento.

Dado que gran parte de las guías están compuestas por textos cortos, sumado a que en ellas

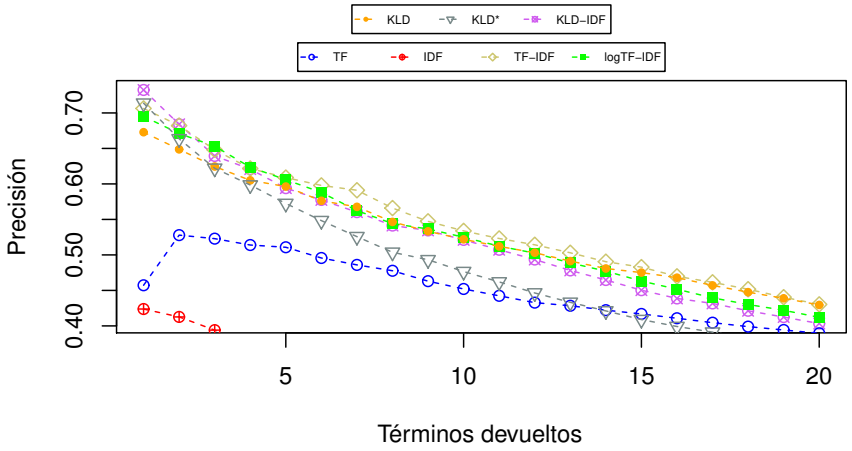


Figura 4.9: Precisión a 20 términos extraídos

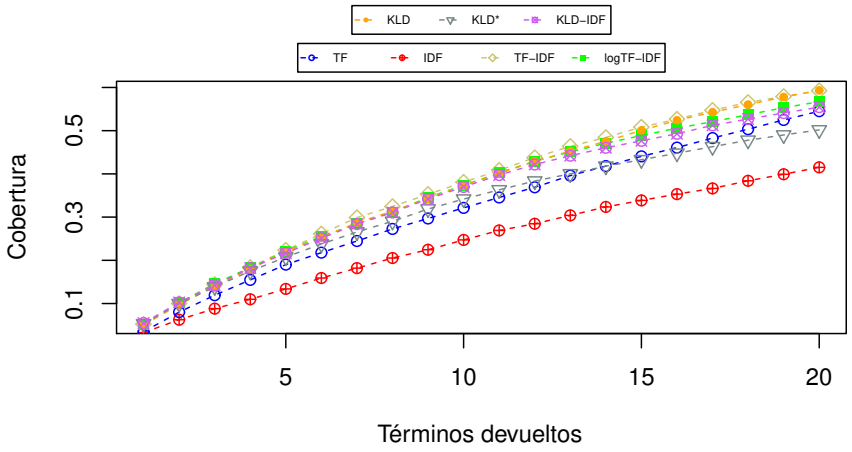


Figura 4.10: Cobertura a 20 términos extraídos

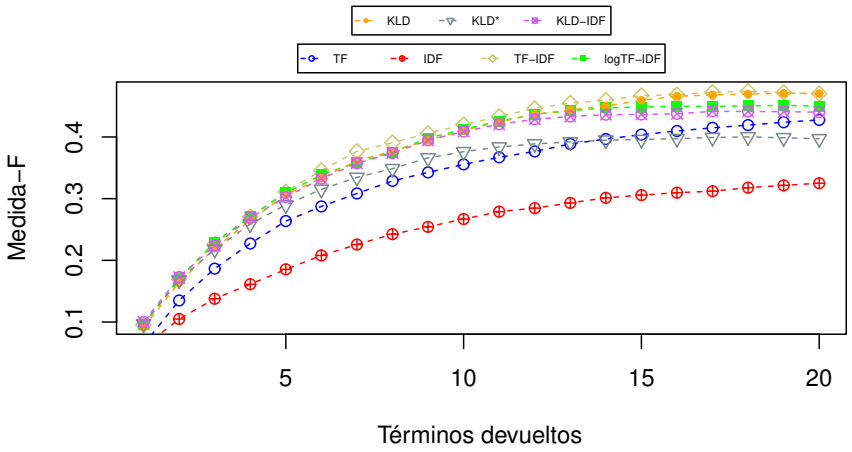


Figura 4.11: Medida-F a 20 armónica términos extraídos

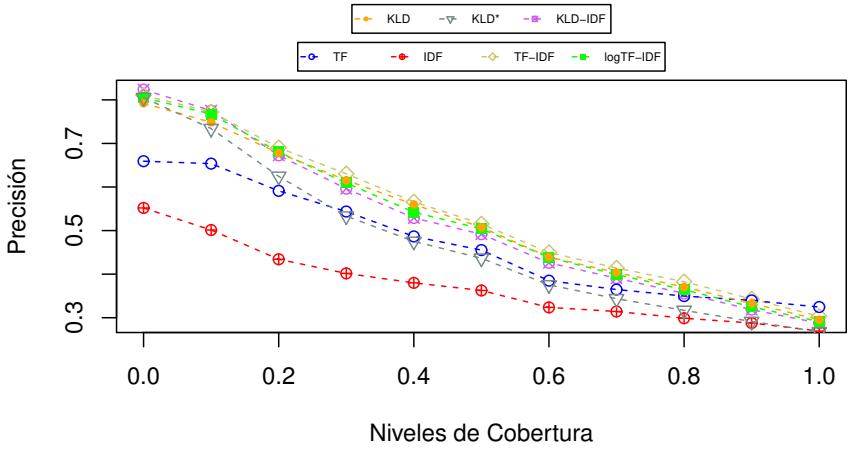


Figura 4.12: Precisión a 11 niveles de cobertura

	MAP	Precisión media a R
TF	0.4509	0.4284
IDF	0.3604	0.3251
TF-IDF	**0.5156**	**0.4875**
LogTF-IDF	0.5021	0.4741
KLD	0.5030	0.4795
KLD*	0.4524	0.4254
KLD-IDF	0.4951	0.4653

Tabla 4.7: Comparación de las funciones para MAP y Precisión a R

los términos suelen aparecer con una frecuencia media de entre una o dos apariciones, según [50] TF se puede utilizar de una manera bastante satisfactoria. Sin embargo, además de estas características, las guías se componen de términos estructurales y términos característicos del dominio, que suelen aparecer repetidos con gran asiduidad dentro de éstas. Este último hecho parece ser el causante de las bajas precisiones que obtiene TF respecto al resto (sobre todo entre los primeros términos extraídos, debido con gran probabilidad a la frecuencia con la que suelen aparecer los términos estructurales en las guías).

Si revisamos de nuevo el apartado 3.2.3 vemos que IDF proporcionará el mismo peso para gran cantidad de los términos del corpus (un 65.5 % de los términos aparecían únicamente en una guía de la colección). Además, veíamos que sólo un 20 % de los términos aparecían en más de 3 guías, y un 4 % se distribuía entre 25 y 1500 guías. IDF, por sí sóla, no ofrece buenos resultados, pero mejora considerablemente si es combinada con TF.

Otra métrica que sostiene los resultados expuestos hasta el momento es la precisión a 11 niveles de cobertura, representada en la gráfica 4.12. En dicha gráfica se puede observar que los mejores resultados son obtenidos por las funciones KLD, KLD-IDF, logTF-IDF y TF-IDF, siendo TF-IDF ligeramente más precisa que el resto. Y de manera adicional, en la tabla 4.7 se presentan las medidas MAP y Precisión Media a Cobertura con resultados análogos a lo presentado hasta el momento.

Hasta ahora se ha debatido qué función utilizar para realizar el poblado de la ontología con el campo *keywords*. Pero existe además otra cuestión muy a tener en cuenta en este punto: ¿cuántos términos utilizaremos para poblar cada asignatura?. En la gráfica 4.11 de medida-F se observa que para las funciones con mejores resultados, el punto de inflexión en la gráfica se

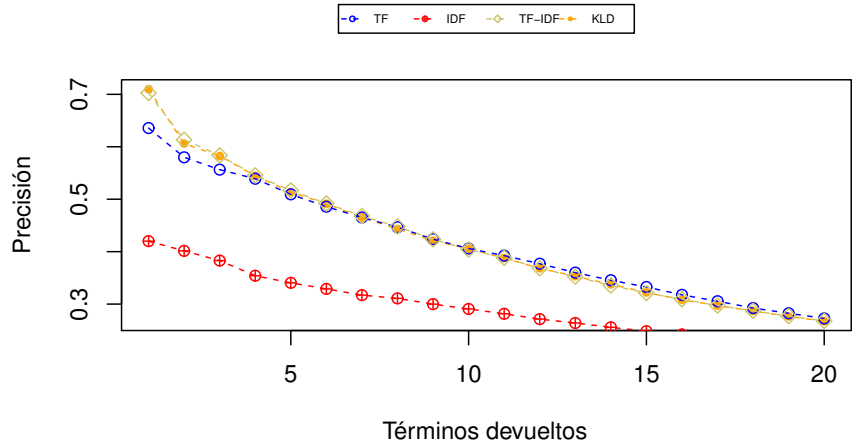


Figura 4.13: Precisión a 20 términos extraídos en corpus libre de términos comunes

encuentra en torno los 18 términos extraídos, lo que nos indica que si queremos que el poblado terminológico tenga un equilibrio entre precisión y cobertura, debemos utilizar en torno a 18 términos por asignatura (podemos coger 20 por que la cifra sea más redonda). Sin embargo, si lo que queremos es maximizar la precisión, podemos tomar entre 1 y 5 términos, con el fin de tener precisiones más altas (entre 1 y 8 para TF-IDF), basándonos en la gráfica 4.9. El enfoque a tomar vendrá determinado por la aplicación final que se le quiera dar a los términos extraídos teniendo en cuenta, si para conseguir el objetivo se necesita maximizar la cantidad de términos importantes o minimizar el número de términos despreciables entre dichos términos extraídos.

Resultados obtenidos por las funciones clásicas sobre el corpus libre de términos comunes

Dadas las características del corpus donde una pequeña parte de los términos presentaban valores extremos en cuanto a frecuencia de aparición en la colección o su distribución a lo largo de ésta, se propuso como alternativa la limpieza de aquellos términos que aparecían al mismo tiempo entre los más frecuentes del corpus (4 % representado en el último cuantil de la tabla 3.7) y entre los que en mayor cantidad de guías se distribuyen (4 % representado en el último cuantil de la tabla 3.9). Estos términos se pueden categorizar como propios de la estructura de la guía o comunes a grandes áreas de conocimiento.

La intersección formada por los grupos de términos citados, y que ha sido utilizada como listado de términos a limpiar del corpus de guías, está compuesta por un total de 1495 términos, entre los cuales encontramos algunos como: asignatura, tema, contenido, estudio, concepto, programa, introducción, análisis, sistema, bloque, desarrollo, general, básico, proceso, social, relación, siguiente, modelo, temático, forma, tipo, aplicación, estructura, investigación, problema, principal, estudiar, aspecto, conocimiento, función, objetivo, teórico, derecho, social, político...

Una vez realizada la limpieza de dichos términos, se prosigue con el pesado del corpus resultante mediante las funciones clásicas presentadas en este trabajo (TF, IDF, TF-IDF y KLD). En la figuras 4.13, 4.14, 4.15 se muestran, respectivamente, las precisiones, coberturas y medidas-f armónicas para estas funciones, pudiendo apreciarse que exceptuando IDF, dichas funciones tienen unos resultados muy similares pero claramente peores que los obtenidos mediante el pesado del corpus completo. Dicho hecho también se manifiesta por los resultados reflejados mediante la figura 4.16 de precisión a 11 niveles de cobertura y la tabla 4.8 con los valores MAP y Precisión Media a Cobertura.

Dados los resultados, se puede deducir que entre los términos más comunes y distribuidos del

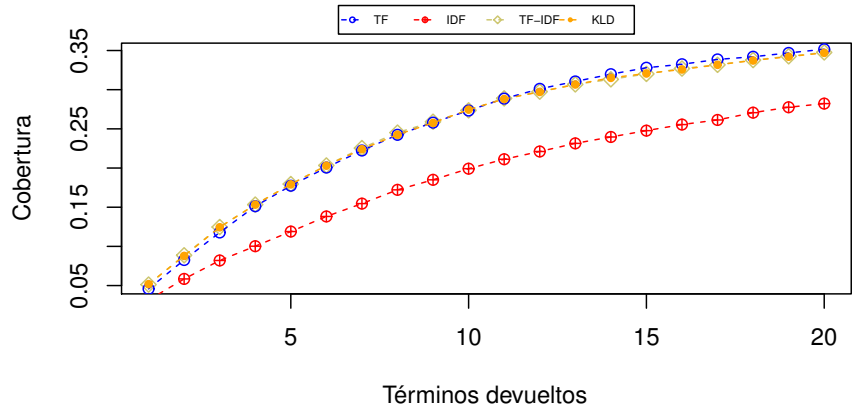


Figura 4.14: Cobertura a 20 términos extraídos en corpus libre de términos comunes

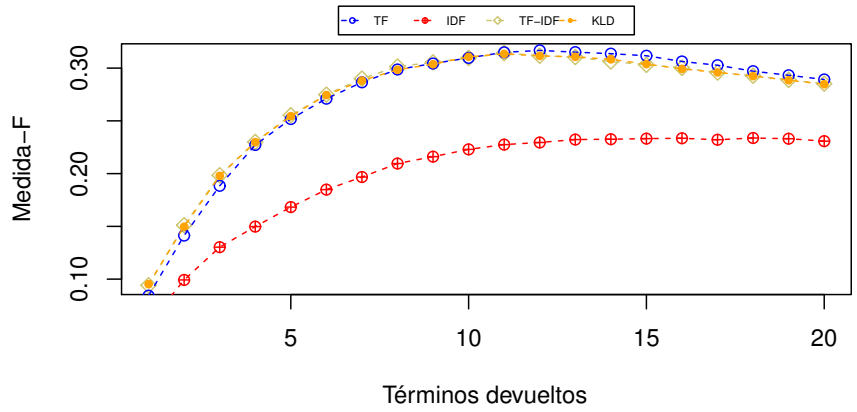


Figura 4.15: Medida-F armónica a 20 términos extraídos en corpus libre de términos comunes

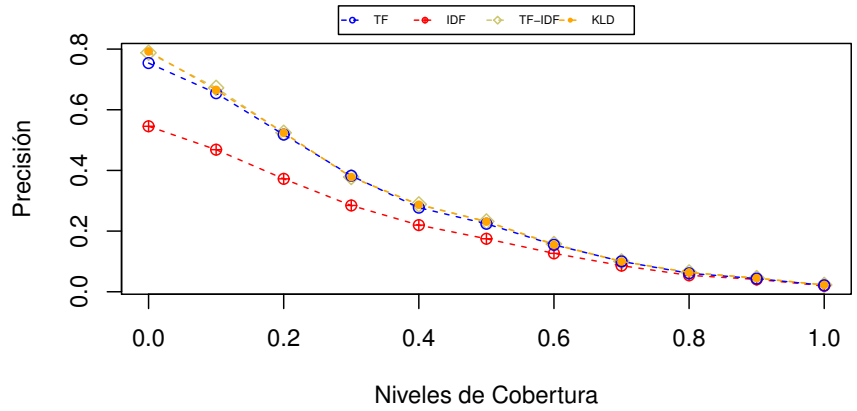


Figura 4.16: Precisión a 11 niveles de cobertura en corpus libre de términos comunes

	MAP	Precisión media a R
TF	0.2662	0.3260
IDF	0.2010	0.2523
TF-IDF	**0.2733**	**0.3280**
KLD	0.2726	0.3274

Tabla 4.8: Comparación de las funciones para MAP y Precisión a R

	Corpus completo							Corpus libre de términos comunes			
	TF	IDF	TFIDF	logTFIDF	KLD	KLD*	KLDIDF	TF	IDF	TFIDF	KLD
p@1	0.4572	0.4238	0.7063	0.6952	0.6729	0.7138	**0.7323**	0.6357	0.4201	0.7026	0.71
p@2	0.5279	0.4126	0.6822	0.671	0.6487	0.6636	**0.684**	0.5799	0.4015	0.6134	0.6059
p@3	0.5229	0.3941	0.6481	**0.653**	0.6245	0.6221	0.6394	0.5564	0.3829	0.5836	0.5824
p@4	0.5139	0.3717	0.6217	**0.6236**	0.605	0.5985	0.6208	0.539	0.3541	0.5455	0.5446
p@5	0.5108	0.3643	**0.6089**	0.6059	0.5963	0.5725	0.5941	0.5093	0.3405	0.5167	0.5138
p@6	0.4957	0.3587	**0.5979**	0.588	0.5762	0.5483	0.5774	0.4857	0.329	0.4919	0.4913
p@7	0.4859	0.3521	**0.5911**	0.5624	0.5677	0.5258	0.5603	0.4652	0.317	0.4679	0.4652
p@8	0.4777	0.348	**0.566**	0.5441	0.5465	0.5037	0.5414	0.4466	0.3109	0.448	0.4433
p@9	0.463	0.3424	**0.5473**	0.5366	0.5333	0.4932	0.5345	0.4238	0.2999	0.4234	0.4209
p@10	0.452	0.3372	**0.5338**	0.5249	0.5219	0.4766	0.5212	0.4059	0.2907	0.4041	0.4056
p@11	0.4424	0.3342	**0.5231**	0.5123	0.512	0.462	0.5073	0.3924	0.2815	0.3883	0.3876
p@12	0.4328	0.3268	**0.5143**	0.5025	0.5031	0.4464	0.4935	0.377	0.2714	0.3683	0.3683
p@13	0.4284	0.3234	**0.503**	0.4887	0.4913	0.4332	0.4781	0.36	0.2639	0.3523	0.3532
p@14	0.4222	0.3213	**0.4904**	0.4772	0.4809	0.4211	0.4647	0.3455	0.2557	0.3354	0.3375
p@15	0.4169	0.316	**0.4828**	0.4632	0.4751	0.4089	0.4503	0.3326	0.2483	0.3212	0.3222
p@16	0.4108	0.3118	**0.4698**	0.4517	0.4679	0.3992	0.4387	0.3176	0.2421	0.3092	0.3083
p@17	0.4045	0.3059	**0.4614**	0.4395	0.457	0.3912	0.4317	0.3055	0.2344	0.2972	0.2972
p@18	0.399	0.304	**0.4517**	0.4302	0.4475	0.3829	0.4213	0.2924	0.2301	0.2873	0.2869
p@19	0.3942	0.3007	**0.4404**	0.422	0.4389	0.3731	0.4124	0.2825	0.224	0.277	0.2765
p@20	0.3896	0.2976	**0.4303**	0.4121	0.4296	0.3638	0.4033	0.2727	0.2173	0.268	0.2673

Tabla 4.9: Comparativa de precisiones

corpus se encuentran algunos que han sido seleccionados como importantes en el *Gold Standard*, pudiendo ser al mismo tiempo comunes a, por ejemplo, un área más global de conocimiento y al mismo tiempo a una determinada guía.

Selección de términos en base a los resultados

Nuestro objetivo es poblar el campo *"keywords"* con el mayor número posible de términos representativos de cada asignatura, evitando en la medida de lo posible que aparezcan otros tipos de términos. Es por esta razón que para el poblado vamos a enfocarnos en la precisión como métrica determinante para este trabajo, y dado que en las precisiones representadas en las figuras 4.9 y 4.13 las curvas que describen algunas de las funciones son bastante cercanas entre sí, presentamos la tabla 4.9 comparativa, donde se muestran las precisiones para los 20 primeros términos extraídos con las distintas funciones empleadas sobre el corpus completo y sobre el corpus libre de aquellos términos más comunes.

La tabla 4.9 muestra que las precisiones obtenidas son mejores, por lo general, para las funciones que han operado sobre el corpus completo de términos, y dentro de este conjunto de resultados podemos considerar TF-IDF como la función más precisa, dado que es la que mejores resultados ofrece desde p@5 (precisión a 5 términos) hasta p@20 términos, quedando de p@1 a p@4 con resultados cercanos aunque inferiores a las funciones ajustadas logTF-IDF (para p@3 y p@4) y KLD-IDF (para p@1 y p@2).

Dicha tabla revela que las funciones logTF-IDF y KLD-IDF devuelven resultados similares

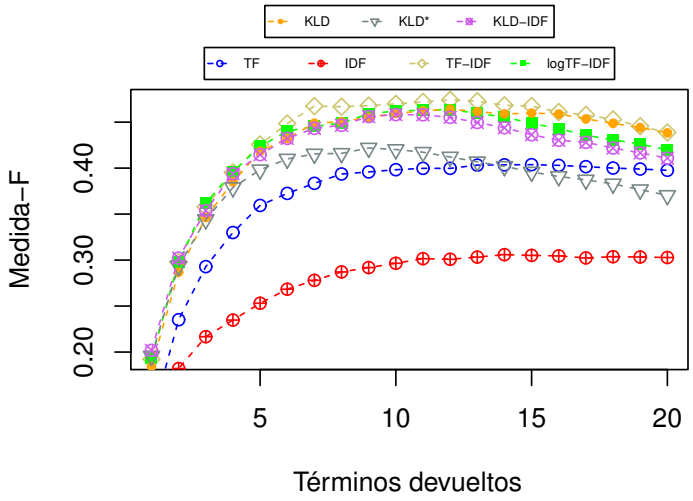


Figura 4.17: Medida-F a 20 términos extraídos para $\beta = 0.5$

aunque ligeramente inferiores a los ofrecidos por TF-IDF (a partir de p@4 y p@2 respectivamente). El objetivo de estas funciones era la devaluación de los términos estructurales de las guías y, sin embargo, parece que pueden haber menospreciado otros términos de cierta importancia. Por ejemplo, las citadas funciones pueden haber penalizado términos comunes a un determinado área de conocimiento (como "derecho" o "social") que resultasen importantes a su vez para la descripción de algunas guías. Así que, TF-IDF a pesar de recuperar en ocasiones términos estructurales que las funciones ajustadas desprecian con toda probabilidad, también recuperará términos que dichas funciones penalizan y que pueden ser característicos de las guías además de ser ciertamente comunes en la colección.

Dados los resultados, en cuanto a precisión se refiere, se ha tomado la decisión de seleccionar los siete términos con mayor peso obtenido mediante la utilización de la función TF-IDF sobre el corpus completo, con el fin de llevar a cabo el poblado terminológico del campo *"keywords"* de la clase "Asignatura" de nuestra ontología.

Otra métrica que podríamos haber utilizado para la toma de la decisión, más allá de puramente la precisión, es la medida-f con un valor β que diera mayor importancia a precisión que a cobertura. En la gráfica 4.17 se muestra la medida-f con $\beta = 0.5$ para los primeros 20 términos recuperados por las distintas funciones sobre el corpus completo. Las conclusiones obtenidas serían las mismas que se han obtenido hasta ahora.

Aunque por exigencias temporales no ha podido realizarse, lo ideal hubiera sido llevar a cabo un análisis de significancia estadística entre las funciones con resultados similares (KLD, TF-IDF, KLD-IDF, logTF-IDF), tarea que se deja pendiente para futuros trabajos.

4.2.3. Resultados del enriquecimiento del *datastore* con palabras clave

Tras llevar a cabo la selección terminológica en cada guía y su poblado con Jena (ver código 3.3), se ha producido un enriquecimiento del *datastore* inicial, lo que conlleva un aumento del potencial tanto en los posibles casos de uso como en nuevos enriquecimientos del mismo. El número de tripletas contenido tras este proceso ha crecido hasta las 125868 (un total de 25251 tripletas nuevas con el enriquecimiento).

A modo de ilustración, se van a presentar tres ejemplos de consulta SPARQL y sus correspondientes resultados hacia el *endpoint* enriquecido:

1. Un alumno desea cursar un Máster en el que pueda ampliar sus conocimientos sobre la

materia de “lógica”. Para ello podría realizar una consulta SPARQL como la que se muestra en el código 4.3, obteniendo los resultados presentados en la tabla 4.10.

Código 4.3: Consulta 1 por términos al SPARQL Endpoint

```
143 PREFIX  :      <http://localhost/ontouned/ontologia/uned#>
144
145 SELECT distinct?NombreMaster
146 WHERE {
147         ?asignatura :nombreAsig ?NombreAsignatura.
148         ?asignatura :keyword ?palabra.
149         ?asignatura :seEnmarca ?estudios.
150         ?estudios :tipoDeEstudios "MASTER".
151         ?estudios :nombreEst ?NombreMaster
152         FILTER regex(str(?palabra), "^logica$")
153 }
```

NombreMaster
"MASTER UNIVERSITARIO EN FILOSOFÍA TEÓRICA Y PRÁCTICA"
"MASTER UNIVERSITARIO EN I.A. AVANZADA: FUNDAMENTOS,MÉTODOS Y APLICACIONES"
"MASTER UNIVERSITARIO EN INVESTIGACIÓN EN INGENIERÍA DE SOFTWARE Y SISTEMAS INFORMÁTICOS"

Tabla 4.10: Resultados a la consulta por términos 1

2. Un coordinador quiere saber qué asignaturas de Grado comparten el estudio de los grafos dentro de sus contenidos principales. Para lograrlo podría realizar una consulta SPARQL como la que se muestra en el código 4.4, obteniendo los resultados presentados en la tabla 4.11.

Código 4.4: Consulta 2 por términos al SPARQL Endpoint

```
157
158 PREFIX  :      <http://localhost/ontouned/ontologia/uned#>
159
160 SELECT distinct?CodigoAsignatura ?NombreAsignatura ?CodigoGrado ?
       Keyword
161 WHERE {
162 ?asignatura :codigoAsig ?CodigoAsignatura .
163 ?asignatura :nombreAsig ?NombreAsignatura .
164 ?asignatura :keyword ?Keyword .
165 ?asignatura :seEnmarca ?estudios .
166 ?estudios :tipoDeEstudios "GRADO".
167 ?estudios :codigoEst ?CodigoGrado
168 FILTER regex ( str(?Keyword ) , "grafo")
169 }
```

CodigoAsignatura	NombreAsignatura	CodigoGrado	Keyword
61021051	MATEMÁTICA DISCRETA	6102	digrafos
61021051	MATEMÁTICA DISCRETA	6102	teoría de grafos
61021051	MATEMÁTICA DISCRETA	7101	digrafos
61021051	MATEMÁTICA DISCRETA	7101	teoría de grafos
71024079	MODELOS PROBABILISTAS Y ANÁLISIS DE DECISIONES	7101	grafo
71901037	LÓGICA Y ESTRUCTURAS DISCRETAS	7101	grafo
71902019	PROGRAMACIÓN Y ESTRUCTURAS DE DATOS AVANZADAS	7101	grafo
61021051	MATEMÁTICA DISCRETA	7102	digrafos
61021051	MATEMÁTICA DISCRETA	7102	teoría de grafos
71024079	MODELOS PROBABILISTAS Y ANÁLISIS DE DECISIONES	7102	grafo
71901037	LÓGICA Y ESTRUCTURAS DISCRETAS	7102	grafo
71902019	PROGRAMACIÓN Y ESTRUCTURAS DE DATOS AVANZADAS	7102	grafo

Tabla 4.11: Resultados a la consulta por términos 2

3. Un docente desea conocer cuáles son las diez palabras clave más comunes en el "GRADO EN INGENIERÍA INFORMÁTICA". Con el fin de obtener esta información podría realizar una consulta SPARQL como la que se muestra en el código 4.5, obteniendo como resultados los presentados en la tabla 4.12.

Código 4.5: Consulta 3 por términos al SPARQL Endpoint

```
172
173 PREFIX  :      <http://localhost/ontouned/ontologia/uned#>
174
175 SELECT (count(?termino) as ?Repeticiones) (?termino as ?Termino)
176 WHERE {
177         ?asignatura :keyword ?termino.
178           ?asignatura :seEnmarca ?estudios.
179           ?estudios :nombreEst "GRADO EN INGENIERIA INFORMATICA"
180 }
181 GROUP BY ?termino
182 ORDER BY DESC(?repeticiones)
183 LIMIT 10
```

Repeticiones	Termino
6	sistema
3	algoritmo
3	diseño
3	grafo
3	planificación
3	programación
3	red
3	software
3	unidad
2	aleatorio

Tabla 4.12: Resultados a la consulta por términos 3

Capítulo 5

Conclusiones y Trabajos Futuros

5.1. Conclusiones

Las conclusiones extraídas de la elaboración de este trabajo se presentan en dos vertientes:

1. Con el desarrollo de este trabajo, en la sección 3.1 hemos visto que se puede producir una transferencia tan directa como se desee desde las bases de datos relacionales a un *datastore* de tripletas RDF de la UNED. Adicionalmente, dicha transferencia se puede realizar de forma periódica bajo demanda o diseñando procesos automáticos de consultas sobre la marcha a las bases de datos relacionales [18]. Además, no existe una necesidad imperiosa de modelar los datos mediante una ontología, por lo que la creación del *datastore* podría ser inmediata dado un acceso autorizado a las bases de datos y respetando las relaciones que establezcan la estructuras de éstas. Sin embargo, es una buena práctica dotar a los datos de una mayor semántica, de manera que puedan establecerse nuevas relaciones entre ellos y así facilitar la labor a los usuarios y agentes que requieran de su consulta o inferencia. Se puede ir aún más allá si utilizamos vocabularios externos como los descritos en la sección 3.3, para el enriquecimiento y externalización de nuestra ontología.

   Además, en la sección 3.2, hemos estudiado cómo se puede enriquecer el *datastore* extrayendo información de recursos referenciados por éste. En este trabajo hemos accedido y extraído información desde fuentes desestructuradas para incorporarla con gran sencillez al *datastore* inicial, produciendo su enriquecimiento. La capacidad para realizar un poblado automático a partir de texto libre lleva asociada una implicación que puede ser exportable a diferentes áreas e instituciones.

   Tras la elaboración del *datastore* y su posterior enriquecimiento, se establece, consecuentemente, un punto donde usuarios y agentes automáticos pueden realizar consultas, modificar e insertar nueva información.

2. La experiencia obtenida a través de la exploración del corpus de guías y de las distintas funciones de pesado. En la sección 2.2 presentamos las funciones KLD y TF-IDF con las que se pesaría el corpus de guías obtenidas durante el trabajo. El objetivo era extraer aquellos términos más específicos de cada guía, siendo en el caso del KLD los términos que aparecieran con mayor probabilidad en la guía y con menor probabilidad en el corpus.

   Sin embargo, tras el estudio del análisis del corpus (en la sección 3.2.3) se apreciaba un corpus de guías bastante especial, donde la gran mayoría de las guías presentaban pocos términos y los términos aparecían con frecuencias muy bajas tanto en la colección como en las guías específicas. Concluimos que esto, sumado a que en la colección de guías aparecían con frecuencia términos estructurales y otros característicos de las distintas áreas de conocimiento, enturbiaría la extracción de términos que únicamente fueran característicos

de la guía. Para solucionarlo, en la sección 3.2.3 se propuso llevar a cabo una metodología que permitiría descartar los términos estructurales de las guías.

Por un lado, se propuso la eliminación directa de determinados términos del corpus, que dadas sus altas frecuencias supondrían una baja importancia específica dentro de sus guías. Para ello se tomó un pequeño porcentaje de términos del corpus que aparecían al mismo tiempo con una alta distribución en las guías y con una alta frecuencia en el corpus de éstas.

Por otro lado, se propuso una modificación de las funciones clásicas KLD y TF-IDF, de tal modo que aquellos términos estructurales y generales quedaran penalizados en relación al resto. Tal y como veíamos en la sección 3.2.3, dada la existencia de términos en algunas guías que aparecen con gran desviación típica en su frecuencia relativa respecto al resto, TF podía aportar un valor al peso que IDF no podría contrarrestar (aún con un gran valor). Para solventarlo, se propone utilizar una versión de TF-IDF que ya se había utilizado en la literatura: logTF-IDF. Por otro lado, KLD padecía del mismo problema que TF-IDF: la probabilidad de aparición de un término en la guía puede ser mucho mayor en términos relativos que la del término en el corpus (recordamos que tenemos un corpus con gran cantidad de términos diferentes y con guías que son, por lo general, cortas), y su KLD presentará un peso alto a pesar de que se tratara del término con mayor número de apariciones de toda la colección. Para solucionarlo se presentan dos modificaciones: penalizar con IDF la función KLD y dar otro enfoque a la función KLD, con KLD*, para que la divergencia de probabilidades tomara mayor relevancia.

Sin embargo, tras la generación de un *Gold Standard* (ver sección 4.2.1) se observó que mediante las métricas estudiadas en la sección 4.2.1, los resultados presentados en la sección 4.2.2, sorprendentemente, favorecían a TF-IDF sobre el corpus completo de términos.

Con el análisis del *Gold Standard* en la sección 4.2.1 se vio que muchas asignaturas incluían como términos importantes, escogidos por los Equipos Docentes, algunos que podían ser más representativos de un área de conocimiento que de una guía en particular (como “social” o “derecho”), y en algunos casos se marcaron como importantes algunos términos estructurales como “tema” (en cinco guías) o “capítulo” (en dos guías).

El hecho de que términos característicos de un dominio de conocimiento sean también importantes dentro de algunas guías puede ser el causante de que TF-IDF obtenga los mejores resultados. Si, por otro lado, descartaramos esta clase de términos, es posible que las funciones ajustadas (KLD*, KLD-IDF y logTF-IDF) pudieran haber obtenido mejores resultados que las clásicas sobre este corpus de guías.

La extracción, aún así, ha sido bastante buena y cercana entre las distintas funciones (ver sección 4.2.2), lo que indica que la existencia de estos términos en el corpus y en el *Gold Standard* no ha influido tanto como se esperaba en la extracción. Quizá se hubiera mejorado ligeramente la obtención del Gold Standard si se hubiera hecho especial hincapié en que no se seleccionaran este tipo de términos. Sin embargo, y por condicionar lo menos posible la selección por parte de los docentes, se optó por no incidir en este tipo de recomendaciones.

Por último, hubiera sido conveniente realizar un análisis de significancia estadística sobre las funciones que ofrecen mejores resultados con el fin de comprobar si los resultados obtenidos por TF-IDF son realmente mejores que los ofrecidos por el resto de las funciones. Sin embargo, debido a las limitaciones temporales que presenta el TFM, este estudio se deja pendiente como futuro trabajo.

5.2. Trabajos futuros

La incorporación de las Tecnologías Semánticas y Datos Enlazados en la UNED conllevan una serie de consecuencias de las que la institución puede beneficiarse, tal y como lo llevan haciendo desde hace tiempo otras Universidades como la Open University de Reino Unido. Para este fin, se propone una serie de líneas y trabajos que pueden desarrollarse a partir del presente.

1. El trabajo inmediato que se deriva del presentado en esta memoria es la publicación del *datastore* como proyecto piloto, con el fin de que la información y los datos contenidos en éste queden disponibles a profesores, usuarios y otros agentes que puedan realizar inserciones y consultas, además de establecer juicios sobre posibles mejoras a realizar. Además, con la discusión abierta en la sección 3.3, se puede iniciar el procedimiento para la redefinición de este vocabulario de partida en términos de ontologías externas más reconocidas.

   Por otro lado, se puede llevar a cabo el establecimiento de un lenguaje común a varias instituciones (tras llevar a cabo una discusión pausada con éstas), que puede realizarse con distintos fines como, por ejemplo: disponer de una mayor facilidad de consulta entre universidades; establecer un sistema comparativo en cuanto a estudios o recursos empleados por las distintas instituciones en el desarrollo de sus funciones; establecer comparativas entre publicaciones realizadas por diferentes grupos de trabajo, localizando investigaciones o investigadores especializados en distintos dominios; etc.

2. Enriquecimiento del *datastore* con nuevas fuentes de datos estructuradas. Por un lado, con toda la información de carácter público que queda por modelar, y por otro con toda la información interna de la Universidad, la cual puede seguir siendo privada pero dotada de semántica para su explotación e inferencia por parte de la UNED. La información disponible interna puede ser conectada de manera controlada con otra información pública, ofreciendo distintos niveles de consulta dependiendo de la publicidad con que se desee dotar a los datos. Por ejemplo, un profesor puede disponer al mismo tiempo de datos privados y públicos dotados de semántica, que permitan cierto nivel de inferencia y consulta por parte de la propia Universidad (o por un departamento, facultad, etc.), y un nivel más limitado de consulta hacia la parte externa a ésta o hacia otros usuarios internos con menos privilegios.

3. Enriquecimiento del *datastore* desde fuentes de datos desestructuradas. La UNED dispone de una gran variedad y cantidad de contenidos textuales, audiovisuales, etc., que pueden servir de ayuda tanto a docentes como estudiantes. Una de las líneas propuestas discurre por la asignación de contenidos audiovisuales recomendados por asignatura mediante por ejemplo, el *matching* entre los *keywords* asignados a las asignaturas y los términos más importantes de los materiales a recomendar (llevando a cabo un análisis y extracción mediante las funciones empleadas en este trabajo u otras según convenga en cada caso).

   Como línea adicional, se propone la búsqueda de una posible mejora de los resultados obtenidos en este trabajo para la extracción terminológica de las guías. Para ello se propone:

   - Encontrar nuevo material y más extenso con el que enriquecer el corpus de guías, de manera que la frecuencia media terminológica no se acerque a 1, para que funciones como KLD realicen su trabajo con mayor facilidad.
   - Llevar a cabo un análisis de significancia estadística de las funciones empleadas en este trabajo.
   - Disponemos de información terminológica a varios niveles de profundidad, de manera que un término no sólo pertenece a una guía, sino que también los cuatrimestres donde se enmarca esta asignatura, los estudios donde aparecen, la rama de conocimiento donde se encuadra, y por último al corpus completo de guías. Se puede establecer un

sistema de pesado con bonificaciones/penalizaciones en función de la globalidad del término. De esta manera, se considerará que los términos irán ganando importancia en función del grado de aparición en los distintos niveles comentados.

4. Dotar de semántica a las palabras clave identificadas durante este trabajo para cada asignatura, de manera que se puedan establecer conexiones y relaciones de distinto tipo entre estos términos mediante un vocabulario controlado como SKOS. De este modo, bajo un mismo concepto pueden existir otros conceptos subclase de éste, llegando a crear taxonomías de conceptos tan complejas como requiera el dominio a explorar.

5. Creación de interfaces para ver la progresión conceptual en un programa de estudios y otras aplicaciones de usuario como las citadas en el trabajo, donde un coordinador pueda establecer comparativas y comprobar si existen deficiencias o solapamientos entre los contenidos de las asignaturas, o también por ejemplo, que un alumno pueda encontrar determinados cursos, estudios o materiales según sus intereses, capacidades, expediente académico, etc.

Aunque se han citado las líneas de trabajo más inmediatas, tras el crecimiento del *datastore* se produciría un incremento de su potencial, pudiendo llegar a establecerse líneas adicionales como los proyectos realizados por otras instituciones, y que han sido descritos en el apartado 2.1.2.

Bibliografía

[1] C. C. Aggarwal and C. Zhai. A survey of text clustering algorithms. In *Mining text data*, pages 77–128. Springer, 2012.

[2] A. Aizawa. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65, 2003.

[3] G. Amati, C. Carpineto, and G. Romano. Comparing weighting models for monolingual information retrieval. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 310–318. Springer, 2003.

[4] T. Berners-Lee. Information management: A proposal. 1989.

[5] T. Berners-Lee. Linked data, 2006, 2006.

[6] T. Berners-Lee, J. Hendler, O. Lassila, et al. The semantic web. *Scientific american*, 284(5):28–37, 2001.

[7] B. Bigi. Using kullback-leibler distance for text categorization. In *European Conference on Information Retrieval*, pages 305–319. Springer, 2003.

[8] I. I. Bittencourt, S. Isotani, E. Costa, and R. Mizoguchi. Research directions on semantic web and education. 2008.

[9] C. Bizer, T. Heath, and T. Berners-Lee. Linked data-the story so far. *Semantic services, interoperability and web applications: emerging concepts*, pages 205–227, 2009.

[10] C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee. Linked data on the web (ldow2008). In *Proceedings of the 17th international conference on World Wide Web*, pages 1265–1266. ACM, 2008.

[11] T. Brants. Natural language processing in information retrieval. In *CLIN*, 2003.

[12] D. Brickley, R. V. Guha, and B. McBride. Rdf vocabulary description language 1.0: Rdf schema. w3c recommendation (2004). *URL http://www. w3. org/tr/2004/rec-rdf-schema-20040210*, 2004.

[13] D. Brickley and L. Miller. Foaf vocabulary specification 0.99. namespace document 14 january 2014-paddington edition. *Retrieved May*, 3:2016, 2014.

[14] P. Brusilovsky et al. Adaptive and intelligent technologies for web-based eduction. *Ki*, 13(4):19–25, 1999.

[15] C. Carpineto, R. De Mori, G. Romano, and B. Bigi. An information-theoretic approach to automatic query expansion. *ACM Transactions on Information Systems (TOIS)*, 19(1):1–27, 2001.

[16] Á. Castellanos, J. Cigarrán, and A. García-Serrano. Generación de un corpus de usuarios basado en divergencias del lenguaje. In *II Congreso Español de Recuperación de Información*, 2012.

[17] W. W. W. Consortium et al. The organization ontology. 2014.

[18] O. Corcho. Ontology-based data integration: State of the art and research challenges, November 2016.

[19] O. Corcho, M. Fernández-López, A. Gómez-Pérez, and A. López-Cima. Construcción de ontologías legales con la metodología methontology y la herramienta webode. 2005.

[20] E. Daga, M. d'Aquin, A. Adamou, and S. Brown. The open university linked data–data. open. ac. uk. *Semantic Web*, 7(2):183–191, 2016.

[21] M. d'Aquin, C. Allocca, and T. Collins. Discou: A flexible discovery engine for open educational resources using semantic indexing and relationship summaries. In *Proceedings of the 2012th International Conference on Posters & Demonstrations Track-Volume 914*, pages 13–16. CEUR-WS. org, 2012.

[22] M. Dean, G. Schreiber, S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, and L. A. Stein. Owl web ontology language reference. *W3C Recommendation February*, 10, 2004.

[23] M. d'Aquin. Linked data for open and distance learning. 2012.

[24] M. d'Aquin. Putting linked data to use in a large higher-education organisation. In *Proceedings of the Interacting with Linked Data (ILD) workshop at Extended Semantic Web Conference (ESWC)*. Citeseer, 2012.

[25] M. Fernández-López, A. Gómez-Pérez, and N. Juristo. Methontology: from ontological art towards ontological engineering. 1997.

[26] C. Galvez, F. de Moya-Anegón, and V. H. Solana. Term conflation methods in information retrieval: non-linguistic and linguistic approaches. *Journal of Documentation*, 61(4):520–547, 2005.

[27] D. Garijo and Y. Gil. Augmenting prov with plans in p-plan: scientific processes as linked data. CEUR Workshop Proceedings, 2012.

[28] Y. Gil, V. Ratnakar, J. Kim, P. Gonzalez-Calero, P. Groth, J. Moody, and E. Deelman. Wings: Intelligent workflow-based design of computational experiments. *IEEE Intelligent Systems*, 26(1):62–72, 2011.

[29] T. R. Gruber et al. A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2):199–220, 1993.

[30] D. Hiemstra. A probabilistic justification for using tf× idf term weighting in information retrieval. *International Journal on Digital Libraries*, 3(2):131–139, 2000.

[31] D. C. M. Initiative et al. Dublin core metadata element set, version 1.1. 2012.

[32] Y. Ji and J. Eisenstein. Discriminative improvements to distributional sentence similarity. In *EMNLP*, pages 891–896, 2013.

[33] M.-R. Koivunen and E. Miller. W3c semantic web activity. *Semantic Web Kick-Off in Finland*, pages 27–44, 2001.

[34] T. G. Kolda. Limited-memory matrix methods with applications. Technical report, 1998.

[35] R. J. Krovetz. Word sense disambiguation for large text databases. 1996.

[36] S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[37] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, and J. Zhao. Prov-o: The prov ontology. *W3C recommendation*, 30, 2013.

[38] H. P. Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, 1(4):309–317, 1957.

[39] A. Maedche and S. Staab. Ontology learning for the semantic web. *IEEE Intelligent systems*, 16(2):72–79, 2001.

[40] C. D. Manning, P. Raghavan, H. Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[41] A. McCallum, K. Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer, 1998.

[42] A. K. McCallumzy and K. Nigamy. Employing em and pool-based active learning for text classification. In *Proc. International Conference on Machine Learning (ICML)*, pages 359–367. Citeseer, 1998.

[43] A. Miles and S. Bechhofer. Skos simple knowledge organization system reference. 2009.

[44] J. Minker, G. A. Wilson, and B. H. Zimmerman. An evaluation of query expansion by the addition of clustered terms for a document retrieval system. *Information Storage and Retrieval*, 8(6):329–348, 1972.

[45] M. A. Musen. The protégé project: A look back and a look forward. *AI matters*, 1(4):4–12, 2015.

[46] K. Oleksiy. Lecture 2: Storing and querying rdf data, Autumn 2016.

[47] L. Padró and E. Stanilovsky. Freeling 3.0: Towards wider multilinguality. In *LREC2012*, 2012.

[48] D. Pinto, J.-M. Benedí, and P. Rosso. Clustering narrow-domain short texts by using the kullback-leibler distance. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 611–622. Springer, 2007.

[49] H. S. Pinto, S. Staab, and C. Tempich. Diligent: Towards a fine-grained methodology for distributed, loosely-controlled and evolving engineering of ontologies. In *Proceedings of the 16th European Conference on Artificial Intelligence*, pages 393–397. IOS Press, 2004.

[50] N. Polettini. The vector space model in information retrieval-term weighting problem. *Entropy*, pages 1–9, 2004.

[51] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[52] M. F. Porter. Snowball: A language for stemming algorithms, 2001.

[53] E. Prud, A. Seaborne, et al. Sparql query language for rdf. 2006.

[54] Y. Qiu and H.-P. Frei. Concept based query expansion. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 160–169. ACM, 1993.

[55] S. Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, 60(5):503–520, 2004.

[56] S. E. Robertson. On term selection for query expansion. *Journal of documentation*, 46(4):359–364, 1990.

[57] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[58] G. Salton and M. J. McGill. Introduction to modern information retrieval. 1986.

[59] G. Salton and M. J. McGill. The smart and sire experimental retrieval systems. In *Readings in information retrieval*, pages 381–399. Morgan Kaufmann Publishers Inc., 1997.

[60] M. Steinbach, G. Karypis, V. Kumar, et al. A comparison of document clustering techniques. In *KDD workshop on text mining*, volume 400, pages 525–526. Boston, 2000.

[61] M. C. Suárez-Figueroa, A. Gómez-Pérez, and M. Fernández-López. The neon methodology for ontology engineering. In *Ontology engineering in a networked world*, pages 9–34. Springer, 2012.

[62] Y. Sure, S. Bloehdorn, P. Haase, J. Hartmann, and D. Oberle. The swrc ontology–semantic web for research communities. In *Portuguese Conference on Artificial Intelligence*, pages 218–231. Springer, 2005.

[63] Y. Sure, C. Tempich, and D. Vrandecic. Ontology engineering methodologies. *Semantic Web Technologies: Trends and Research in Ontology-based Systems*, pages 171–190, 2006.

[64] M. Uschold and M. Gruninger. Ontologies: Principles, methods and applications. *The knowledge engineering review*, 11(02):93–136, 1996.

[65] A. Voutilainen. Part-of-speech tagging. *The Oxford handbook of computational linguistics*, pages 219–232, 2003.

[66] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 4–11. ACM, 1996.

[67] F. Zablith, M. Fernandez, and M. Rowe. The ou linked open data: production and consumption. In *Extended Semantic Web Conference*, pages 35–49. Springer, 2011.

[68] L. Zemmouchi-Ghomari and A. Ghomari. Process of building reference ontology for higher education. In *Proceedings of the World Congress on Engineering*, volume 3, pages 1595–1600, 2013.