

# Búsqueda semántica en bases documentales gubernamentales

Fundación CTIC: Emilio Rubiera, Jose María Álvarez,  
Diego Berrueta, Iván Frade, Luis Polo  
Universidad de Oviedo: Enrique del Teso,  
Roger Bosch, Natalia Cueto, Jose Emilio Labra

26 de febrero de 2007

## Resumen

Este documento tiene como objetivo presentar el buscador *Esperteyu*. Este buscador es un prototipo desarrollado por el Equipo de Web Semántica de la Fundación CTIC, en colaboración con profesorado de la Universidad de Oviedo. El objetivo que trata de resolver el buscador *Esperteyu* se enmarca dentro de las políticas generales de Sociedad de Información del Principado de Asturias para acercar la Administración Pública a los ciudadanos. En particular, el buscador *Esperteyu* intenta mejorar y facilitar el acceso de la ciudadanía al Boletín Oficial del Principado de Asturias (BOPA).

El proyecto aplica distintas tecnologías semánticas, en particular, ontologías y tesauros. La construcción del buscador se ha realizado buscando la máxima interoperabilidad con sistemas existentes utilizando los estándares del W3C: XML, XHTML, CSS, OWL-DL o SKOS-CORE. Se presentarán los fundamentos teóricos más importantes que subyacen al desarrollo del buscador, además de ofrecer una panorámica, muy superficial, de la arquitectura y filosofía que sostienen nuestro modelo particular de búsqueda vertical.

# Índice general

<b>1. Introducción</b>	<b>2</b>
<b>2. Búsqueda semántica y caracterización formal de los documentos</b>	<b>5</b>
2.1. Recuperación de documentos etiquetados con metadatos . . .	7
2.2. Búsqueda sintáctica guiada por los conceptos de una ontología	8
<b>3. Enriquecimiento de las consultas de los usuarios</b>	<b>13</b>
3.1. Formalización del proceso de enriquecimiento semántico . . .	14
3.2. Visión general del enriquecimiento de consultas . . . . .	16
3.3. Validez de la consulta . . . . .	17
3.4. Transformación 0 . . . . .	17
3.5. Transformación 1 . . . . .	17
3.6. Transformación 2 . . . . .	19
3.7. Transformación 3 . . . . .	19
3.8. Ejecución de la búsqueda . . . . .	20
<b>4. Redes semánticas y técnicas de <i>Spreading Activation</i></b>	<b>21</b>
4.1. Definición del algoritmo . . . . .	22
4.2. Parametrización del <i>Spreading Activation</i> . . . . .	23
4.3. Implementación de <i>Spreading Activation</i> . . . . .	24
4.4. Nuestra versión de <i>Spreading Activation</i> . . . . .	26
<b>5. Arquitectura del sistema de conocimiento: ontologías y tesauros</b>	<b>29</b>
5.1. Tesauros . . . . .	31
5.2. Cuestiones preliminares sobre las ontologías . . . . .	33
5.3. <i>Framework</i> semántico . . . . .	34
5.4. Metodología y conceptos generales . . . . .	38
5.5. Organización de las ontologías . . . . .	41

# Capítulo 1

## Introducción

Según la propia definición del W3C “la Web Semántica es una Web extendida, dotada de mayor significado en la que cualquier usuario en Internet podrá encontrar respuestas a sus preguntas de forma más rápida y sencilla gracias a una información mejor definida”. Los esfuerzos por explotar tecnológicamente el modelo RDF de la Web Semántica han desembocado en varias líneas de investigación. La capacidad de representar semánticamente en este modelo las propiedades de documentos constituyen el fundamento para aplicar ontologías a la búsqueda de información. En este “technical report”, nuestra intención es presentar el proyecto BOPA. Este proyecto se enmarca dentro de las iniciativas del Principado de Asturias por mejorar el acceso a la información publicada por la Administración. Su objetivo es construir un buscador semántico para el Boletín Oficial del Principado de Asturias (BOPA), y acercar la Administración Pública a los ciudadanos.

Una persona que esté buscando una información tiene uno o más objetivos en mente, y los utiliza sistemas de búsqueda como herramientas para alcanzar esos objetivos. El modelo de los procesos de acceso a la información asume (véase Figura 1.1)(FIX), independientemente del grado de experiencia del usuario, un ciclo de interacción que consiste en la especificación de una pregunta (técnicamente, una *query*), el examen de los resultados y si estos no satisfacen los objetivos iniciales, reformular la pregunta y repetir el proceso de búsqueda.

Este modelo consiste en una secuencia estructurada de pasos, lo que se conoce técnicamente como un **guión**, y lo hemos asumido como un comportamiento característico en cualquier situación de búsqueda. Uno de los aspectos críticos en el diseño del buscador fue decidir cuál iba a ser la entrada al sistema porque eso definiría la interacción con el usuario, y a la postre, su tipo de comportamiento. Los usuarios de Internet ya estamos muy habituados a los buscadores tipo *Google* o *Yahoo*, y a los interfaces de acceso sencillos mediante el uso de *keywords*. Nuestra forma de buscar en Internet ha modificado en gran parte nuestra forma de interactuar con los sistemas de

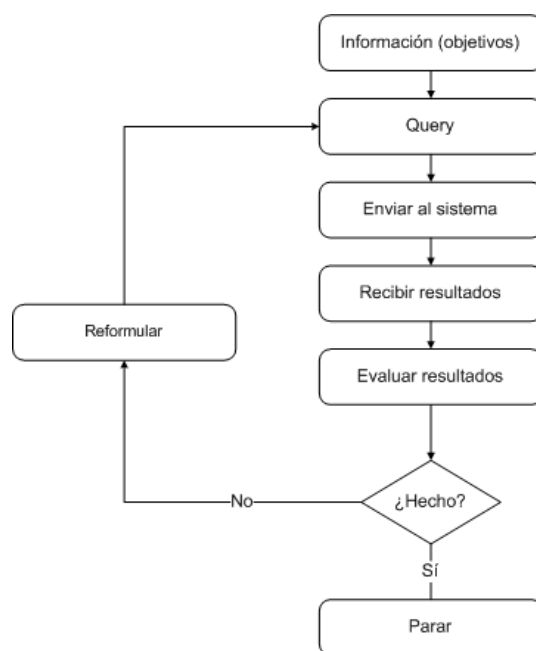


Figura 1.1: Modelo estándar de los procesos de acceso a la información

búsqueda, porque básicamente el procedimiento es cómodo e intuitivo. Finalmente, decidimos una interacción simple para ofrecer al usuario la mayor usabilidad posible, continuando la metodología de los buscadores tradicionales sintácticos. El *input* para nuestro sistema de búsqueda semántica es una cadena de texto compuesta por las palabras (o términos) que el usuario escoge como representativas para alcanzar sus objetivos de búsqueda.

A la hora de afrontar el diseño del buscador semántico, nos pareció importante plantearnos cuáles son o pueden ser los objetivos potenciales de búsqueda de los usuarios y qué esperan encontrar en una base documental gubernamental, como es el caso de un Boletín Oficial. Este tipo de análisis y el estudio experto de los contenidos que se publican en el Boletín son las pistas fundamentales para entender qué clase de semántica es la adecuada para representar la información.

Las administraciones públicas generan cada día gran cantidad de información de consumo público, como son ofertas de empleo público, convocatorias de ayudas y subvenciones, notificaciones de distinto tipo, etc. Esta información, se publica en España en los boletines oficiales del Estado, de las Comunidades Autónomas y de sus respectivas provincias. Además de esta información de tipo administrativo, también se publican en estos boletines toda la normativa con alcance dentro del territorio correspondiente: leyes, reglamentos, planes generales de ordenación urbana, etc. El BOPA es una de estas colecciones de textos oficiales. En resumen, podemos asumir

que los documentos del BOPA pueden ser o bien normas reguladoras o bien disposiciones sobre actos concretos de la Administración.

Otra de las cuestiones críticas ha sido el léxico que se utiliza en los documentos oficiales, y que está muy alejado del vocabulario del ciudadano medio. Son textos escritos con una terminología muy técnica, tanto legal como administrativa. Éste ha sido uno de los mayores problemas que nos hemos encontrado a la de abordar la construcción del buscador semántico, un escollo muy habitual en el caso de las búsquedas verticales, donde la terminología del dominio suele ser muy específica y concreta. Este problema del léxico no es relevante para los usuarios expertos, como juristas, funcionarios, y otro tipo de personal cualificado. La barrera léxica aparece del lado del ciudadano cuando intenta localizar los artículos. Palabras como “cohecho”, “alícuota” o “predio” son bastante inaccesibles y están lo suficientemente alejadas de su léxico habitual como para que no sean incluidas en ninguna consulta. De esta forma, los términos de búsqueda de los usuarios rara vez encajan con el vocabulario técnico y específico del legislador y con la corrección política obligatoria en cualquier documento oficial. Un ejemplo claro en este sentido es el término “persona mayor de 70 años” que difícilmente será utilizado como término de consulta por ningún ciudadano.

En resumen, detectamos como mínimo tres problemas, que se deben solucionar, si aspiramos a construir una plataforma de búsqueda semántica que satisfaga razonablemente las expectativas de los usuarios.

1. Analizar la estructura del BOPA y la tipología de documentos que la constituyen.
2. Superar la barrera léxica entre el vocabulario del BOPA y el léxico de entrada al sistema de los usuarios.
3. Añadir conocimiento al motor de búsqueda para completar la falta de información de los ciudadanos acerca de la estructura de la administración pública y del sistema normativo y legislativo.

## Capítulo 2

# Búsqueda semántica y caracterización formal de los documentos

Los motores de búsqueda sintácticos (basados en la búsqueda de cadenas de texto coincidentes) han demostrado en la última década tener una capacidad increíble para localizar y recuperar información con precisión a lo largo y ancho de los contenidos que pueblan la Web. El uso de algoritmos como “PageRank”<sup>1</sup>, en el caso de Google, para puntuar la importancia de un documento combinado con el refinamiento de las técnicas estadísticas de búsqueda de texto han aumentado la potencia y precisión de búsqueda en la Web. Por eso, hay que tener bien presente que las tecnologías semánticas no tratan de suplantar la búsqueda sintáctica, sino de ofrecer soluciones en dominios donde el uso de este tipo de técnicas no ofrecen el rendimiento esperado.

Cuanto más vertical es el espacio de búsqueda, y más técnica y específica es la terminología del dominio, más complicado tiene un usuario ajeno a ese dominio de formular consultas con criterios adecuados para que los motores de búsqueda ofrezcan resultados satisfactorios. Además, en el caso del BO-PA, Google no podría aplicar su “PageRank” para construir una puntuación del documento a partir de los links entre las páginas, ya que no existen este tipo de links y las relaciones entre documentos, por ejemplo, un reglamento que haga referencia a una ley determinada, son conceptuales y no pueden ser interpretadas por un buscador sintáctico. Nosotros consideramos que el uso

---

<sup>1</sup>Google “PageRank” descansa en la propia naturaleza de la Web. Google utiliza las referencias entre páginas como un indicador del valor de un determinado documento web. Así, Google interpreta un link desde una página  $A$  a una página  $B$  como un voto de  $A$  a  $B$ . Pero además del volumen de votos totales de una página web,  $B$ , Google analiza la página de la que recibe el voto,  $A$ . De forma que cuanto más importante es  $A$ , mayor puntuación recibe la página  $B$ .

de las ontologías, como mecanismos formales para representar conocimiento de un dominio concreto, pueden ayudar a un sistema de búsqueda a focalizar las consultas de los usuarios y a llegar a sitios donde el puro “matching” sintáctico entre cadenas de caracteres no es suficiente.

Nuestra propuesta se basa en la aplicación de las nuevas tecnologías semánticas para la recuperación de documentos del BOPA. Nuestro sistema combina dos procesos de búsqueda complementarios, en el que cada uno explota determinadas propiedades de los textos. En primer lugar, un acercamiento ortodoxo de la Web Semántica (empleo de metadatos como método de descripción de propiedades de los documentos) y un enfoque híbrido que, mediante la aplicación de ontologías y tesauros, permite enriquecer semánticamente la consulta del usuario y ejecutar búsquedas más precisas. Para el usuario esto es invisible, ya que sólo puede acceder a la base documental mediante el cajetín único de consulta. Es el sistema el que debe reconocer qué información debe utilizar para cada tipo de procedimiento.

¿Por qué necesitamos dos tipos de búsqueda? Los documentos de un Boletín Oficial, llamados técnicamente “disposiciones”, son un tipo de objeto social conocido como objeto informativo<sup>2</sup>. Los objetos informativos son recursos (entidades) que estructuran información. Tipos de objeto informativo pueden ser libros, carnés de identidad, carteles de publicidad, etc. En el caso de documentos y textos en general, los objetos informativos tiene determinadas propiedades *per se*:

1. Propiedades extensionales: estas propiedades son independientes del contenido del texto. La fecha de publicación, el autor o un identificador numérico, como es el caso de un ISBN o la clave de una biblioteca.
2. Propiedades intensionales: el contenido de un texto. Un texto construye un discurso, un conjunto de oraciones que constituyen una unidad de significado.

La formalización y la explotación de las propiedades de los documentos es bastante distinta según sea el caso. Mientras que para el tratamiento de las propiedades extensionales<sup>3</sup> existen procedimientos, vocabularios de descripción y herramientas de gestión y control, en el caso de las propiedades intensionales es completamente distinto. Y es que, aunque la teoría y el análisis del discurso han alcanzado su madurez en los últimos años, la representación semántica de un texto es bastante compleja. No se conocen

---

<sup>2</sup>Para la realización del análisis de los documentos del BOPA, hemos utilizado la propuesta de ontología fundacional DOLCE.

<sup>3</sup>Las propiedades extensionales son la base de las herramientas de gestión documental utilizadas en biblioteconomía y archivística, así como el alcance de vocabularios de descripción de recursos, como por ejemplo, el estándar Dublin Core (DC).

herramientas de representación del conocimiento que sean lo suficientemente flexibles y expresivas para este campo que sean completas.<sup>4</sup>

## 2.1. Recuperación de documentos etiquetados con metadatos

El concepto de *metadato*<sup>5</sup> precede a Internet y a la Web Semántica, sin embargo, el interés en el desarrollo de estándares de metadatos y aplicaciones capaces de explotarlos ha crecido exponencialmente con la publicación electrónica y el problema de la “sobrecarga de información” en la red. Los metadatos son una herramienta básica para la organización y clasificación de documentos, pero, sobre todo, para la recuperación de información. A partir de los metadatos podemos filtrar un subconjunto determinado de documentos, ayudando en muchos casos a los usuarios que no son capaces de formular eficazmente una consulta. En nuestro caso, hemos considerado como metadatos de los documentos del Boletín sus propiedades extensionales.

Hemos organizado los metadatos de los documentos del Boletín bajo la perspectiva de un sistema de clasificación facetado. Informalmente podemos expresar esto como una taxonomía de taxonomías. Técnicamente, en nuestro caso, nuestro sistema de clasificación se organiza en una ontología en la que los contenidos (los documentos BOPA) se describen a través de varias dimensiones o **facet**as. Estas son estructuras lógicas de conceptos y constituyen cada uno de los aspectos relevantes a la hora de describir las disposiciones.

Para desarrollar nuestra clasificación facetada, hemos reutilizado vocabularios diseñados con el propósito exclusivo de tratar los objetos informativos: propiedades del vocabulario Dublin Core: *dc:creator*, *dc:identifier*, *dc:date*, y propiedades de la ontología fundacional DOLCE: *edns:about* y *edns:expresses*. Estas propiedades permiten describir el autor de una disposición del BOPA, la fecha, utilizar su identificador numérico y definir los tipos o categorías textuales que existen en el BOPA.

Hemos construido una ontología utilizando el lenguaje OWL-DL para representar y formalizar estos metadatos. Esta ontología modela la estructura básica del Boletín:

---

<sup>4</sup>Problemas de esta índole llevan planteados desde hace décadas en el campo de la Traducción Automática. La búsqueda de formalismos y herramientas para representar y gestionar adecuadamente el contenido de un texto es una de sus principales preocupaciones.

<sup>5</sup>Información relativa a otra información.

$$\text{DocumentoBOPA} \sqsubseteq \exists dc : creator. (\text{ÓrganoAdministrativo}) \quad (2.1)$$

$$\text{DocumentoBOPA} \sqsubseteq \exists dc : date. (\text{xsd:date}) \quad (2.2)$$

$$\text{DocumentoBOPA} \sqsubseteq \exists dc : identifier. (\text{xsd:integer}) \quad (2.3)$$

$$\text{DocumentoBOPA} \sqsubseteq \exists edns : about. (\text{Norma} \sqcup \text{ActoAdministrativo}) \quad (2.4)$$

$$\text{DocumentoBOPA} \sqsubseteq \forall edns : about. (\text{Norma} \sqcup \text{ActoAdministrativo}) \quad (2.5)$$

Tanto el tipo de categoría como el emisor del artículo permiten clasificar los textos y crear clases disjuntas de documentos recuperables a partir de los metadatos de los propios textos.

El tipo de categoría es especialmente relevante en esta cuestión. La categoría de una disposición depende de a qué haga referencia el propio documento. La propia estructura del Boletín hace que un texto sólo pueda referirse (*edns:about*) a dos tipos de entidades:

- Actos de la Administración Pública: convocatoria o resolución de subvenciones, notificación a un particular, resolución de un procedimiento de licitación, etc.
- Normas: como pueden ser leyes, reglamentos, ordenanzas municipales, etc.

En este sentido, si un ciudadano está buscando ayudas o subvenciones para material escolar, un buscador no debería proporcionarle disposiciones de convocatorias de licitación para la construcción de escuelas o colegios públicos. No sólo es necesario identificar el contexto administrativo en el que se ubican las consultas de los usuarios, sino también el tipo de disposiciones que se adecúan a sus objetivos.

## 2.2. Búsqueda sintáctica guiada por los conceptos de una ontología

El caso de las propiedades intensionales es diferente. Desde el principio, intuíamos que existía un cierto vacío en el uso de metadatos para la definición de recursos. El problema: cómo etiquetar el contenido de los artículos. Los metadatos son una forma de gestión documental (bibliotecas, grandes organizaciones como la administración pública, archivos, etc.) muy explorada. Sin embargo, tiene sus límites y el principal escollo es cómo tratar el discurso textual y convertirlo en un criterio de búsqueda. El uso de propiedades como “materia” o “tópico”<sup>6</sup> es lo más habitual. Los vocabularios

---

<sup>6</sup>En el mismo sentido que puede tener la a propiedad *dc:subject* del vocabulario Dublin Core.

controlados, como taxonomías o tesauros, se utilizan para la indexación de los documentos y para describir la “materia” del recurso mediante palabras clave o códigos de descripción de una clasificación terminológica. En el Principado de Asturias, el Servicio de Documentación y Archivos se encarga de anotar la base documental con términos del tesoro oficial Eurovoc. Cada disposición se etiqueta con dos o tres términos. Sin embargo, ¿qué ciudadano conoce el lexico de Eurovoc, un tesoro de más de 7,000 términos? ¿tiene el Eurovoc el alcance y la profundidad semántica necesaria para abarcar todo el vocabulario de cada dominio de la Administración Pública?<sup>7</sup>

El uso de metadatos para el etiquetado del contenido de los textos presenta varios problemas: 1) tienen un coste de mantenimiento muy alto debido a su dependencia del sistema de representación que se haya utilizado, 2) cualquier cambio en el sistema puede implicar un cambio en el etiquetado, y 3) utilizar vocabularios controlados, como el caso de Eurovoc, no garantizan una mejora real de las búsquedas. También surgieron varios interrogantes técnicos sobre las características del etiquetado: ¿manual?, ¿realizado por un experto<sup>8</sup>?, ¿semiautomático?, ¿podíamos diseñar un procedimiento de confianza para etiquetar automáticamente la intensidad de un documento?, ¿qué profundidad debía tener el etiquetado y los sistemas de clasificación que usásemos?, etc. En conclusión, el uso de metadatos nos pareció insuficiente para el tratamiento y definición del significado de los textos<sup>9</sup>.

Entonces, ¿cómo tratamos las propiedades intensionales? En nuestro sistema de búsqueda, en vez de intentar representar semánticamente el contenido de los artículos, hemos formalizado y estructurado mediante ontologías las competencias y ámbitos de la Administración Pública. Hemos organizado conceptualmente cada parcela administrativa y legislativa, proporcionando un modelo genérico que se puede utilizar como representación, no de los textos en sí, si no de la actividad, procedimientos, estructura y organización de la Administración Pública y sus instituciones. Los documentos del Boletín articulan en su discurso semánticas de dominio relativas a determinados aspectos del mundo: leyes sobre ruido y accesibilidad, reglamentos para el uso de instalaciones deportivas, presupuestos sobre el próximo ejercicio del gobierno, subvenciones para la rehabilitación de fachadas, etc. En este sentido, entendemos que los documentos oficiales del BOPA articulan una determinada descripción del mundo real, que se corresponde con una de nuestras ontologías. Cada una de estas ontologías de dominio  $O_{\mathbb{D}_i}$  organiza un conjunto de conceptos relevantes que representamos como  $\mathbb{D}_i$ . Por eso

---

<sup>7</sup>Este tipo de sistema también se usa en la gestión de bases documentales como las bibliotecas, donde suele existir una ordenación por materias de los distintos textos.

<sup>8</sup>Esta opción quedó descartada desde el inicio del proyecto. Ni somos expertos de dominio, ni podemos plantearnos etiquetar más de 10,000 disposiciones hacia atrás en el tiempo.

<sup>9</sup>El uso de otras técnicas de representación de conocimiento como las que se emplean en campos como NLP o AI estaban fuera del alcance del proyecto.

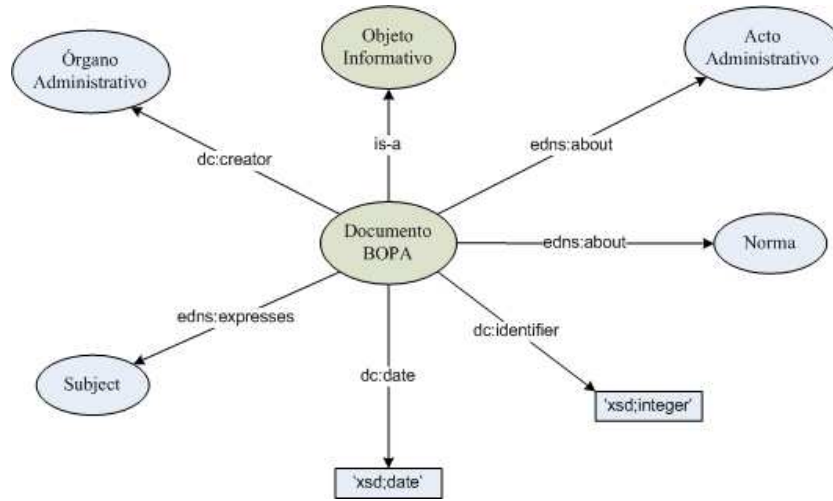


Figura 2.1: Modelo formal de un documento del Boletín Oficial del Principado de Asturias

mismo, nos hemos visto obligados a dividir el sistema de conocimiento del buscador.

1. Ontología del Boletín Oficial del Principado de Asturias.
2. Ontologías que formalizan la semántica de un dominio de la Administración Pública:  $O_{\mathbb{D}_i}$ .

Hemos desarrollado un mecanismo para poder utilizar las ontologías de dominio como valores de descripción de los documentos del Boletín. Para ello es necesario mantener las ontologías de dominio por separado de las ontología que formaliza la estructura del BOPA (así se pueden utilizar con fines concretos que detallaremos más adelante) y a la vez reificarlas dentro de esta ontología para que los documentos del Boletín puedan ser descritos a través de ellas. Cada ontología es capaz de representar la intensión de un conjunto de documento de la Administración. De esta forma, conseguimos tratar la intensión como si fuese una propiedad extensional más, cómo lo haríamos mediante el uso de sistemas de clasificación como los tesauros, pero con la diferencia que en nuestro caso, el valor que se escoge para representar el contenido del documento es una ontología determinada  $O_{\mathbb{D}_i}$ . Una ontología que si hacemos “zoom” en su estructura, descubrimos la semántica compleja y estructurada de un dominio. Para adecuar nuestro tratamiento a los métodos de clasificación documental con vocabularios contralados, consideramos el concepto *Subject* como rango de valores que expresa un documento (*edns:expresses*), y donde  $O_{\mathbb{D}_i} \in \text{Subject}$  (véase Figura 2.1):

$$\text{DocumentoBOPA} \sqsubseteq \forall \text{expresses. (Subject)} \quad (2.6)$$

Pensemos ahora en el ciudadano que se decide a consultar el Boletín Oficial de su provincia. Como señalamos al principio de este documento, el usuario de un buscador tiene en mente una serie de objetivos que desea satisfacer: quiere obtener una determinada información. Si aislamos a este usuario en una situación concreta de búsqueda, el estado mental de este individuo sólo tiene activos una serie de conceptos relevantes para este propósito. Nuestra hipótesis, que desarrollaremos formalmente en la sección siguiente, es que el usuario expresa una consulta sintáctica ( $\alpha$ ) formada por una serie de términos que encapsula un conjunto de conceptos  $\mathbb{C}_\alpha$  de su estado mental.

En el caso de que exista una intersección entre el conjunto de conceptos que configuran los objetivos del usuario,  $\mathbb{C}_\alpha$ , y el conjunto de conceptos que representan un dominio concreto,  $\mathbb{D}_i$ , ese dominio es **relevante** para el usuario y, por extensión, también cualquier documento que esté en relación *edns:expresses* con la ontología que formalice ese dominio  $\mathbb{D}_i$ .

El diseño de nuestro buscador aborda la anterior hipótesis de la siguiente manera: dado un criterio de búsqueda cualquiera  $\alpha$  de un usuario (que refleja una serie de conceptos configurantes de sus objetivos de búsqueda  $\mathbb{C}_\alpha$ ), existe una función  $\wp_{\mathbb{D}}$  que identifica qué dominio  $\mathbb{D}_i$  se corresponde con  $\alpha$  y una función  $\psi_{rel}$  que asigna una puntuación<sup>10</sup> a los documentos que se relacionan con ese dominio (y, por tanto, con los objetivos del usuario):

$$\begin{aligned} \wp_{\mathbb{D}}(\alpha) = \mathbb{D}_i \wedge (\exists x \text{DocumentoBOPA}(x) \wedge \text{expresses}(x, \mathbb{D}_i)) \Rightarrow & \quad (2.7) \\ \psi_{rel}(x, \mathbb{D}_x) = \text{score} \wedge (\text{score} > 0) & \end{aligned}$$

La cuestión ahora es cómo se definen e implementan las funciones  $\wp_{\mathbb{D}}$  y  $\psi_{rel}$ . Estas operaciones implican la definición del propio proceso de búsqueda y cómo sus elementos forman parte del mismo, es decir, en qué sentido las ontologías, por ejemplo, contribuyen a la búsqueda. Necesitamos por tanto determinar cómo se identifican los conceptos a partir de la consulta del usuario<sup>11</sup> ( $\wp_{\mathbb{D}}$ ) y cómo se utilizan los conceptos de la ontología de dominio para localizar y recuperar artículos del Boletín ( $\psi_{rel}$ ). La solución que hemos adoptado, que veremos más en detalle en la siguiente sección, consiste en enriquecer la consulta del usuario y transformar su *query* inicial en una nueva *query* construida a partir de los conceptos de la ontología de dominio seleccionada. El sistema utiliza dos tecnologías semánticas complementarias:

---

<sup>10</sup>Grado de relevancia del documento.

<sup>11</sup>proyecto de Iván

1. Ontologías: para el control y organización de los conceptos del dominio a partir de un modelo de conocimiento (transformación semántica de la *query* inicial)
2. Tesoros: para el control y organización de los términos del dominio y del vocabulario común (identificación de los conceptos subyacentes a la *query* inicial del usuario y ejecución de la búsqueda sintáctica a partir de la transformación semántica de la *query* inicial).

Resumiendo, hemos formalizado las propiedades extensionales e intensionales de cada artículo del BOPA de forma que podemos por un lado, aplicar los metadatos como criterio de búsqueda extensional y, por otro, hemos conseguido utilizar, como se verá más adelante, búsqueda sintáctica dirigida por la semántica de una ontología como método de aproximación entre el significado de los documentos y los objetivos de búsqueda de los usuarios. Hemos tenido que integrar este proceso de búsqueda dirigido semánticamente por una ontología como si fuese una propiedad extensional más, reificando la ontología dentro de la estructura de descripción de los documentos. Los dos procedimientos de búsqueda, metadatos y búsqueda dirigida, corren independientemente hasta la integración de sus resultados, por lo que la activación de determinados elementos de un dominio desencadenará, por un lado, la construcción de una nueva *query* con la semántica de una ontología, y, por otro, la identificación de una ontología como valor de la propiedad *edns:expresses*. A partir de aquí, utilizando el lenguaje de consultas SPARQL, podemos extraer qué tipo de documento es y qué posibles organismos de la Administración han podido emitir disposiciones al respecto, ya que el fondo cada ontología modela una parcela del ámbito administrativo y jurídico, lo que determina tanto qué disposiciones se pueden emitir y qué organismos tienen capacidad competencial en ese dominio. El ejemplo más claro es Urbanismo, donde un tipo de documento específico del dominio son los Planes Generales de Ordenamiento Urbano, que son competencia de los Ayuntamientos de cada concejo.

## Capítulo 3

# Enriquecimiento de las consultas de los usuarios

El sistema de búsqueda es híbrido, combina técnicas semánticas con la tradicional búsqueda sintáctica. Durante el proceso se aplican automáticamente una serie de transformaciones a la consulta del usuario, convirtiendo una **query** sintáctica en una representación semántica equivalente, a partir de las relaciones entre los conceptos de una ontología y el conocimiento lingüístico recogido en los tesauros. Las ontologías modelan varios dominios relacionados con la actividad de la Administración Pública o con aspectos de la legislación. Una pregunta de un usuario está compuesta, desde un punto de vista psicológico, por un conjunto de conceptos y un dominio al que pertenecen estos conceptos. La interfaz de usuario, como dijimos, no es más compleja que la de un motor de búsqueda convencional. Tanto los conceptos como los dominios a los que pertenecen están ocultos al usuario.

Los pasos a realizar son los siguientes:

1. El usuario introduce los términos de búsqueda
2. A partir de los términos de entrada del usuario, se activan los conceptos subyacentes en el sistema de ontologías utilizando y explotando el conocimiento lingüístico y léxico de los tesauros como índice de correspondencias. (transformación 1)
3. El conjunto de conceptos activados en una ontología es el *input* para las técnicas basadas en *Spreading Activation*. Este algoritmo, como se verá en profundidad más adelante, explota la riqueza semántica de la base de conocimiento esencialmente como un explorador de grafos. Dado un conjunto inicial de nodos (los conjuntos activados por la consulta del usuario), el algoritmo recorre los arcos del grafo de la ontología, activando los nodos que están relacionados más estrechamente.

Los conceptos de las ontologías reciben una puntuación de forma dinámica. El conjunto de conceptos inicial que se corresponde con la consulta del usuario recibe la máxima puntuación. Posteriormente, el algoritmo de *Spreading Activation* puntúa sólo los conceptos que estén relacionados semánticamente. Cuanto más lejos estemos de los conceptos iniciales, más baja será la puntuación que reciba ese nodo, ese concepto. El algoritmo se detiene cuando la puntuación es demasiado baja o cuando no quedan más conceptos por explorar. El *output* es una lista de conceptos clasificados por relevancia.(transformación 2)

4. La lista de conceptos es transformada en una lista de palabras, inversamente al procedimiento del primer paso. Cada concepto tiene asociado por la estructura básica del tesoro un *synset*<sup>1</sup> compuesto por términos que encajan perfectamente con el vocabulario administrativo y legislativo del BOPA. (transformación 3)
5. Como último paso, se construye una consulta sintáctica y se ejecuta sobre la base de documentos XML con el buscador sintáctico Lucene. Lucene devuelve un conjunto de documentos ordenados por relevancia.

Una vez que se ha aplicado este procedimiento, el ciudadano obtendrá como resultado una serie de artículos en los cuales podrá encontrar tanto términos que haya introducido como otros nuevos, fruto del algoritmo de propagación de conceptos. De esta manera, aunque se realiza una búsqueda sintáctica, la consulta ha sido creada mediante la activación a través de las relaciones de un conjunto de conceptos de un dominio.

### 3.1. Formalización del proceso de enriquecimiento semántico

A continuación definimos formalmente la serie de transformaciones necesarias en cada paso, desde la introducción de los términos de búsqueda de usuario hasta la consulta última generada automáticamente por nuestro sistema.

Es necesario una definición previa de los conceptos siguientes:

- Sea  $Q_{str}$  una cadena de texto introducida por el usuario en una herramienta de búsqueda.

---

<sup>1</sup>Un *synset* identifica de forma única un concepto. Está constituido por un conjunto de términos intercambiables en un contexto determinado. Es decir, está formado por un conjunto de términos sinónimos. De esta forma, el *synset* del concepto #caballo = {“caballo”, “rocín”, “jamelgo”}.

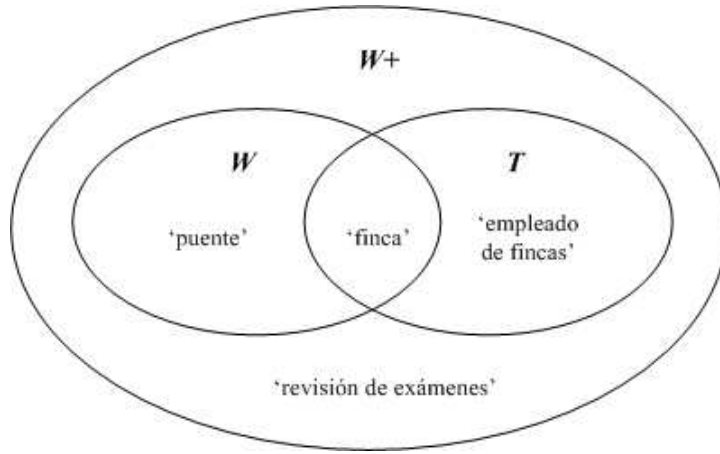


Figura 3.1: Relaciones entre los conjuntos  $W$ ,  $W^+$  y  $T$ .

- Sea  $Q_{sin} = \{t_1, t_2, \dots, t_l\}$  el conjunto de términos que representa la misma información que  $Q_{str}$ .
- Sea  $Q_{sem} = \{c_1, c_2, \dots, c_m\}$  el conjunto de conceptos asociados a los términos anteriores.
- Sea  $Q'_{sem} = \{(c'_1, w_1), (c'_2, w_2), \dots, (c'_n, w_n)\}$  el conjunto de conceptos ponderados que representa la consulta enriquecida. Cada  $w_i$  es un valor numérico en el intervalo  $[0, \infty)$  que indica la relevancia del concepto respecto a la consulta del usuario (a mayor valor, mayor relevancia).
- Sea  $Q'_{sin} = \{(t'_1, w_1), (t'_2, w_2), \dots, (t'_r, w_r)\}$  el conjunto de términos ponderados que se utilizan finalmente en la consulta sintáctica.

**Términos** Sea  $T$  el conjunto de términos de nuestra base de conocimiento léxica.  $T \subset W^+$ , donde  $W$  es el conjunto de todas las palabras de un lengua (el español, por ejemplo) y  $W^+$  es el conjunto de todos los términos posibles (en español).

Por definición,  $W \subset W^+$ . La relación entre  $T$ ,  $W$  y  $W^+$  queda más clara en la Figura 3.1.

**Conceptos** Sea  $C$  el conjunto de conceptos de nuestra base de conocimiento semántica.  $C \subset C^u$ , donde  $C^u$  es el conjunto de todos los conceptos posibles que pueden ser representados por la mente humana.

**Dominios** Sea  $\mathbb{M} = \{\mathbb{D}_1, \mathbb{D}_2, \dots, \mathbb{D}_s\}$  el conjunto de posibles dominios, donde cada dominio<sup>2</sup>  $\mathbb{D}_i \subset \mathbb{C}$ . Nótese que los dominios no son necesariamente disjuntos, es decir, un concepto puede existir en varios dominios.

**Función de interpretación** Sea  $intr : \mathbb{C} \rightarrow \mathbb{D}^+$  la función de “interpretación”, que relaciona cada concepto con los dominios en los que este concepto existe, por lo tanto, está definida como:

$$intr(c_i) = \{\mathbb{D}_j : \forall \mathbb{D}_j / c_i \in \mathbb{D}_j\} \quad (3.1)$$

**Tesoro** El tesoro construido como parte de la base de conocimiento se representa como una relación denominada *lex*. Se trata de una relación entre  $\mathbb{C}$  y  $\mathbb{T}$ , que hace corresponder a cada concepto de  $\mathbb{C}$  los términos de  $\mathbb{T}$  por los que se puede representar. Debido a la sinonimia y la polisemia, no se trata de una función ni de una relación biyectiva. En lo que sigue, se usará también la relación  $lex^{-1}$ .

### 3.2. Visión general del enriquecimiento de consultas

El proceso de enriquecimiento de una búsqueda consiste en la siguiente secuencia de transformaciones:

$$Q_{str} \xrightarrow{0} Q_{sin} \xrightarrow{1} Q_{sem} \xrightarrow{2} Q'_{sem} \xrightarrow{3} Q'_{sin} \quad (3.2)$$

En lo que sigue, se describe cada una de estas transformaciones. Previamente, se expone un ejemplo.

Sea la cadena de texto introducida por el usuario ( $Q_{str}$ ) igual a “vacaciones del empleado de fincas”. La secuencia de transformaciones es la siguiente:

- $Q_{str} =$  “vacaciones del empleado de fincas”
- $Q_{sin} = \{$  “vacacion”, “empleado”, “finca”  $\}$
- $Q_{sem} = \{$  #vacaciones, #empleadodefincas  $\}$
- $Q'_{sem} = \{$  (#vacaciones, 1,0), (#empleadodefincas, 1,0), (#conveniolar, 0,5)  $\}$
- $Q'_{sin} = \{$  (“vacaciones”, 1,0), (“empleado de fincas”, 1,0), (“portero”, 1,0), (“convenio laboral”, 0,5)  $\}$

---

<sup>2</sup>Cada uno de estos dominios  $\mathbb{D}_i$  está formalizado en una ontología correspondiente  $O_{\mathbb{D}_i}$ .

### 3.3. Validez de la consulta

Una consulta es válida si todos los conceptos que forman su expresión semántica tienen exactamente un dominio en común. Es decir:

$$Q_{sem} \text{ es válida} \Rightarrow \text{card}(\bigcap \text{intr}(c_i)) = 1 \quad (3.3)$$

- Si dicha intersección fuese el conjunto vacío, la consulta no sería válida porque no habría un contexto o dominio común.
- Si, por el contrario, dicha intersección tuviese una cardinalidad mayor que uno, la consulta sería ambigua. En este último caso, la situación es recuperable con la intervención del usuario para deshacer la ambigüedad.

Sea  $\mathbb{D}_{com}$ <sup>3</sup> el dominio común a los conceptos de la consulta, y se define como:

$$\{\mathbb{D}_{com}\} = \bigcap \text{intr}(c_i) \quad (3.4)$$

### 3.4. Transformación 0

La primera transformación que sufre la cadena de búsqueda introducida por el usuario es realizada por un conjunto de filtros integrados en el *framework* del motor de búsqueda sintáctica Lucene. Básicamente, consiste en simplificar y normalizar las palabras de la consulta, eliminando palabras no relevantes (*stop words*), singularizando las palabras, unificando la capitalización, etc.

Para representar la transformación llevada a cabo en esta fase, se define la función  $\varphi_0 : \mathbb{W}^+ \rightarrow \mathbb{W}^+$ .

### 3.5. Transformación 1

El objetivo de esta transformación es reconocer los conceptos subyacentes a los términos de búsqueda introducidos por el usuario. La función que realiza esta transformación es  $\varphi_1 : \mathbb{W}^+ \rightarrow \mathbb{C}^+$ , es decir,  $Q_{sem} = \varphi_1(Q_{sin})$ .

Esta función se define recursivamente tomando como base la relación  $lex^{-1}$ :

---

<sup>3</sup>En nuestro sistema,  $\mathbb{D}_{com}$  se corresponde con una y sólo una ontología. El sistema no trabaja con varios dominios a la vez. Esto hace que la división de la actividad administrativa sea una tarea delicada.

$$\varphi_1(\alpha) = \begin{cases} \emptyset & \text{si } \alpha = \lambda \\ \{c : \forall c/(\alpha, c) \in lex^{-1}\} & \text{si } \alpha \in \mathbb{T} \\ \varphi_1(\alpha_1) \cup \varphi_1(\alpha_2) \cup \varphi_1(\alpha_3) & \text{donde } \alpha = \alpha_1 \cdot \alpha_2 \cdot \alpha_3 \\ & \text{tales que } \alpha_2 \in \mathbb{T} \wedge \alpha_1 \cdot \alpha_3 \neq \lambda \\ \perp & \text{en otro caso} \end{cases} \quad (3.5)$$

La tercera alternativa introduce la recursividad en la definición, tratando de “trocear” la cadena  $\alpha$  en subcadenas (estrategia “divide y vencerás”). Obviamente, puede haber varias posibles divisiones de la cadena  $\alpha$ . En ese caso, se opta por la división que maximiza la longitud de  $\alpha_2$ . Cabe señalar que esta elección puede tener consecuencias sobre la posibilidad de evaluar la función (más adelante se examina un ejemplo).

La cuarta alternativa señala que la función no está definida en algunos casos (como es natural, no es posible encontrar la correspondencia en conceptos de cualquier cadena de búsqueda). En estos casos, la búsqueda falla y la consulta debe reformularse.

A continuación, se examinan unos ejemplos. Suponiendo que la relación  $lex^{-1}$  contiene:

$$\begin{array}{lll} \text{“vacacion”} & \xrightarrow{lex^{-1}} & \#vacaciones \\ \text{“piscina”} & \xrightarrow{lex^{-1}} & \#piscina \\ \text{“vigilante”} & \xrightarrow{lex^{-1}} & \#vigilante \\ \text{“vigilante piscina”} & \xrightarrow{lex^{-1}} & \#vigilantepiscina \\ \text{“piscina cubierta”} & \xrightarrow{lex^{-1}} & \#piscinacubierta \end{array} \quad (3.6)$$

Si  $Q_{sin}$  es {“vacacion”, “vigilante”, “piscina”} (que puede haberse derivado de una consulta  $Q_{str}$  = “vacaciones del vigilante de una piscina”), se entra por la tercera rama de la definición, y se escoge  $\alpha_2$  = “vigilante piscina”. Finalmente, el resultado es el conjunto  $Q_{sem} = \{\#vacaciones, \#vigilantepiscina\}$ .

Si, en cambio,  $Q_{sin}$  es {“vigilante”, “piscina”, “cubierta”} (que puede haberse derivado de una consulta  $Q_{str}$  = “vigilantes de piscinas cubiertas”), habría varias posibilidades para la tercera rama:

- Tomar  $\alpha_2$  = “vigilante piscina”, que es la subcadena  $\alpha_2$  más larga posible tal que  $\alpha_2 \in \mathbb{T}$ . Lamentablemente, eso conduce a la evaluación de  $\varphi_1(\alpha_3) = \varphi_1(\text{“cubierta”}) = \perp$ , por lo que la evaluación falla.
- En cambio, si se tomase  $\alpha_2$  = “piscina cubierta”, que no es la subcadena  $\alpha_2$  más larga, la evaluación concluiría satisfactoriamente con el resultado  $Q_{sem} = \{\#vigilante, \#piscina cubierta\}$ .

### 3.6. Transformación 2

Esta transformación consiste, fundamentalmente, en la aplicación de las técnicas de *Spreading Activation*, que se define en la sección 4 (fix) para generar un conjunto ampliado de conceptos en el cual, además, cada concepto tiene asociado un valor numérico que se interpreta como su relevancia en relación a la consulta.

La transformación se formaliza como la función:

$$\varphi_2 : \mathbb{C}^+ \rightarrow (\mathbb{C}, \mathbb{R})^+ \quad (3.7)$$

La transformación convierte el conjunto  $Q_{sem} = \{c_1, c_2, \dots, c_m\}$  en el conjunto  $Q'_{sem} = \{(c'_1, w_1), (c'_2, w_2), \dots, (c'_n, w_n)\}$ . Se representa por *conceptos*( $Q'_{sem}$ ) el conjunto  $\{c'_1, c'_2, \dots, c'_n\}$ .

Se cumple que:

$$Q_{sem} \subseteq \text{conceptos}(Q'_{sem}) \subseteq \mathbb{D}_{com} \quad (3.8)$$

Los valores  $w_i$  que forman el segundo elemento de los pares denotan la relevancia de cada concepto en relación con la consulta del usuario (para cierta definición de “relevancia” que no se comenta aquí FIX). Por tanto,

$$w_i > w_j \Leftrightarrow c_i \text{ es más relevante que } c_j \quad (3.9)$$

Esencialmente, las técnicas de *Spreading Activation* tratan la ontología como un grafo  $G = (V, E)$ , donde el conjunto  $V$  de vértices es  $\mathbb{D}_{com}$  y los arcos dirigidos ( $E$ ) señalan las relaciones entre los conceptos en la ontología. El resultado es un subgrafo  $G' = (V', E')$ , donde  $V' = \text{conceptos}(Q'_{sem})$ . En este subgrafo  $G'$ , se verifica que:

$$c_i \in \text{conceptos}(Q'_{sem}) \Rightarrow \exists \text{ un camino entre } c_j \in Q_{sem} \text{ y } c_i \quad (3.10)$$

### 3.7. Transformación 3

La última transformación vuelve a convertir los conceptos en términos. Para ello, se aplica a cada concepto de  $Q'_{sem}$  la función  $\varphi_3 : (\mathbb{C}, \mathbb{R}) \rightarrow (\mathbb{T}^*, \mathbb{R})$ , que puede considerarse como opuesta de la función  $\varphi_1$  antes descrita. En este caso, la definición es más simple.

$$\varphi_3(c, score) = \{(t, score) : \forall t / (c, t) \in lex\} \quad (3.11)$$

por tanto,

$$Q'_{sin} = \{\varphi_3(c, score) : \forall c, score / (c, score) \in Q'_{sem}\} \quad (3.12)$$

Debido a la sinonimia, pueden existir varios términos para un mismo concepto. Por conveniencia, también pueden existir conceptos sin términos asociados (de ahí que se emplee el cierre estrella). Éste será el caso de conceptos muy abstractos, que juegan un papel instrumental durante la aplicación del *Spreading Activation*, pero no tienen valor sintáctico.

### 3.8. Ejecución de la búsqueda

La consulta  $Q'_{sin}$  tiene el formato adecuado para ser ejecutada por el motor de búsqueda Lucene. Para ello, se concatenan todos los términos ponderados que la componen, utilizando el operador de disyunción. La activación del concepto #portero:  $\{(\#portero, w_i)\}$ , genera la siguiente consulta sintáctica:  $Q'_{sin} = \{("portero", w_i) \text{ OR } ("conserje", w_i) \text{ OR } ("ujier", w_i)\}$

No corresponde a este documento explicar el funcionamiento del motor de búsqueda sintáctica. Es obvio, al utilizar la disyunción, que el número de resultados será potencialmente elevado. No obstante, la ponderación introducida en los términos de búsqueda logrará ordenar los resultados de forma que los más relevantes ocupen las primeras posiciones. Afinar esta ponderación y ajustar el conjunto de conceptos activados y su relevancia es uno de los puntos de desarrollo críticos en los que estamos investigando actualmente.

## Capítulo 4

# Redes semánticas y técnicas de *Spreading Activation*

Las técnicas de *Spreading Activation* nacieron en el campo de la Psicología (fix), como resultado de la investigación de la memoria humana y, más especialmente, en la búsqueda de procedimientos para explotar las formas de representación del conocimiento humano.

El modelo básico de *Spreading Activation* consiste en una red de nodos interconectados, un grafo. Si los nodos representan objetos o clases del dominio y los arcos, relaciones que se establecen entre ellos, podemos hablar entonces de una red semántica. El procesamiento realizado por el algoritmo se basa en un método de exploración de grafos utilizando un modelo iterativo. Cada una de las iteraciones se compone de una serie de pulsos y condición de parada, en los que cada pulso está formado, a su vez, por distintos pasos de ejecución.

La utilización de *Spreading Activation* como algoritmo de exploración de grafos no es nueva y ya a principios de los años 80[59] aparecían los primeros trabajos de investigación nombrando a este algoritmo. Como antecedentes podríamos nombrar trabajos en torno a redes neuronales, redes semánticas y a los algoritmos clásicos de búsqueda en grafos. Su uso se centra principalmente en el campo de “Information Retrieval” y “Document Retrieval” [40]. Aunque con el apabullante éxito de Internet en los últimos años, el uso de este algoritmo ha derivado en obtener documentos de hipertexto[1] o multimedia de la red. Por otra parte, y más parecido al enfoque que pretendemos dar, encontramos proyectos relacionados con un procesos de búsqueda híbridos[63] en los que se intenta generar una consulta sintáctica basada en la exploración previa de una base de conocimiento.

En nuestro buscador, el *Spreading Activation* sólo se aplica a las ontologías de dominio, a los conceptos que constituyen un ámbito particular de la Administración.

## 4.1. Definición del algoritmo

Las técnicas de *Spreading Activation* son un método para explorar redes semánticas a partir de un conjunto inicial de conceptos con determinada puntuación asociada. A partir de este conjunto, la “activación” se propaga iterativamente a los demás conceptos relacionados con ese conjunto inicial. Los “pesos” de los conceptos suelen ser valores reales que decrecen según la activación se propaga por el grafo. Los valores asociados a la entrada de los nodos y los pesos de las relaciones entre los conceptos son configurables dependiendo de la aplicación. A los efectos de que se formen núcleos de activación, los pesos de las relaciones deben ser menores que 1. Evidentemente, esto modificará también su interpretación. La implementación del algoritmo se divide en las siguientes fases,

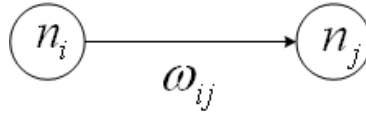


Figura 4.1: Modelo gráfico del *Spreading Activation*

**Ajuste previo (*preadjustement*):** esta fase inicial, de carácter opcional, se suele utilizar para realizar alguna estrategia de control sobre el grafo que se va a explorar.

**Propagación (*spreading*):** fase de expansión del algoritmo. Los conceptos se van activando por oleadas, en las que el nodo tratado activa a sus nodos vecinos.

El cálculo del grado de activación  $I_i$  de un nodo  $n_i$  se realiza mediante la fórmula:

$$I_i = \sum_j O_j \omega_{ji} \quad (4.1)$$

Donde  $I_i$  es el total de entradas del nodo  $n_i$ ,  $O_j$  es la salida del nodo  $n_j$  conectado al nodo  $n_i$  y  $\omega_{ji}$  es el peso de la asociación del nodo  $n_j$  con el nodo  $n_i$ . Si no existe relación entre el nodo  $n_j$  y el nodo  $n_i$  se asume que  $\omega_{ji} = 0$ . Más adelante se discute cómo se calcula el valor de salida  $O_j$  de un nodo  $n_j$  a partir de su valor de entrada  $I_j$ .

Para evaluar el “peso” de un nodo y decidir si un concepto está activo o no, se utiliza una función  $f$  de evaluación de activación definida como:

$$N_i = f(I_i) \quad (4.2)$$

La función de activación  $f$  puede devolver diferentes valores dependiendo nuevamente del ámbito de la aplicación y de la interpretación que queramos asignar a este valor. No obstante, el caso más habitual es considerar como posibles valores  $N_i$ , 1 y 0, que indican si el nodo ha sido activado o no respectivamente:

$$N_i = f(I_i) = \begin{cases} 0 & \text{si } I_i < J_i \\ 1 & \text{si } I_i > J_i \end{cases} \quad (4.3)$$

Denotamos por  $J_i$  el valor de activación umbral para  $i$ . Este valor  $J_i$  es dependiente de la aplicación y es habitual que varíe de un nodo a otro, aunque también puede ser constante. Hay que tener en cuenta que el grado de activación  $I_i$  de un nodo  $n_i$ , a medida que se vaya iterando el algoritmo, puede ir cambiando. Una vez que se satisfaga la regla de terminación, bien por número de iteraciones  $k$  o por otras restricciones establecidas como un valor mínimo de activación, el algoritmo finalizará devolviendo un conjunto de nodos  $\mathcal{G}^k$  y para cada nodo  $n_i$  un peso o grado de activación  $I_i$ , es decir, un conjunto constituido por pares ordenados de la forma  $(n_i, I_i)$ .

**Ajuste posterior (*postadjustment*):** es la fase final. Está destinada al control de los conceptos activados y es opcional.

## 4.2. Parametrización del *Spreading Activation*

El algoritmo de propagación se caracteriza por su flexibilidad desde el punto de vista de la configuración. Por eso, se han establecido una serie de restricciones[18] como mejoras del algoritmo para ajustar los resultados de la activación de conceptos a cada caso de aplicación, y que no se produzcan valores erróneos.

**Distancia:** Los nodos que se encuentren alejados de un nodo activado deberían ser penalizados ya que es necesario realizar muchos saltos para poder llegar a activarlos.

**Definición 4.2.1.** *La distancia del nodo  $n_j$  al nodo  $n_i$ ,  $d_{ji}$ , en una red conceptual es el número mínimo de vértices que deben recorrerse para llegar del nodo  $n_j$  al  $n_i$ .*

**Camino:** El camino seguido por la activación desde un nodo puede ser guiado atendiendo a los pesos y a las etiquetas de las relaciones. Nuevamente, esta característica se configura dependiendo de la aplicación con el objetivo de afinar los resultados de la propagación obtenidos.

**Múltiples salidas (Fan-Out):** Los Nodos “altamente conectados”, con muchas relaciones de salida con otros nodos, pueden desvirtuar el objetivo de la propagación activando nodos que probablemente no se encuentran en el conjunto de activación deseado. Se configura entonces un umbral para determinar el nivel de conectividad máximo permitido para la activación del nodo y su propagación.

**Nota 4.2.2.** *Un nodo  $n_i$  será activado con el peso correspondiente  $I_i$  si su grado de salida,  $g_i$ , el número de arcos que salen de  $n_i$ , es inferior a una constante umbral  $\ell$ .*

**Umbral de activación:** Como comentamos en la fase de *Spreading* de definición del algoritmo, se puede parametrizar la activación de un nodo a partir de un valor umbral mínimo establecido y una función  $f$  de evaluación de la activación. Si el valor de activación de un nodo está por debajo del valor umbral mínimo establecido este nodo no debería ser propagado.

**Nota 4.2.3.** *Un nodo  $n_i$  será propagado si su valor de activación,  $I_i$ , es mayor que una constante umbral de activación  $j$ .*

La aplicación de estas restricciones simplifica y restringe el algoritmo de *Spreading Activation*, permitiendo la adaptación a problemas concretos y afinando los valores de activación de los conceptos.

### 4.3. Implementación de *Spreading Activation*

Nuestro buscador utiliza las técnicas de *Spreading Activation* para, a partir de los conceptos de la consulta inicial del usuario ( $Q_{sem}$ ), extraer un nuevo conjunto de conceptos relevantes ( $Q'_{sem}$ ) que permitan construir una “*query* enriquecida” con terminología de dominio ( $Q'_{sin}$ ). Los conceptos están organizados en un sistema de ontologías, donde cada ontología  $O_i$  formaliza y modela un dominio determinado  $\mathbb{D}_i$ .

Las ontologías son sistemas de representación de conocimiento que, estructuralmente, podemos contemplar como una **red semántica**, por tanto, como un grafo donde cada nodo  $n_i$  representa un concepto  $c_i$  de la ontología de dominio y el arco  $\omega_{ji}$  la relación semántica entre los conceptos  $c_j$  y  $c_i$ . La terminación del algoritmo ofrece como resultado el conjunto de pares ordenados  $(n_i, I_i)$  que forman  $Q'_{sem}$ , donde  $n_i \approx c_i$  y  $I_i \approx w_i$ , la relevancia del concepto.

La arquitectura de ontologías propone un modelo en el cual los conceptos pueden pertenecer a varios dominios, en la que el valor semántico de un concepto (las propiedades que lo definen) puede variar de un dominio a otro. Por ello, la consulta de un usuario es válida si todos los conceptos pertenecen al mismo dominio  $\mathbb{D}_{com}$ . Así el lanzamiento del algoritmo de activación de conceptos sólo se realiza en el dominio  $\mathbb{D}_{com}$  seleccionado a partir de la consulta del usuario. El objetivo de esta restricción será obtener conceptos relevantes en el ámbito que estamos buscando, de esta forma el conjunto de conceptos de salida es siempre semánticamente correcto. Nos encontramos ante una condición necesaria de activación y propagación.

**Definición 4.3.1.** *Sea  $\mathbb{D}_{com}$  el dominio activo para una búsqueda, si un concepto  $c_i$  es activado y propagado, entonces  $c_i \in \mathbb{D}_{com}$ .*

El objetivo de utilizar estas técnicas de activación es seleccionar un conjunto adecuado de conceptos para la localización de documentos cercanos a los intereses de los ciudadanos que consultan el BOPA. La exploración de la ontología como una red semántica nos permite extraer a partir de la consulta del usuario (que refleja en algún sentido sus objetivos de búsqueda) conceptos cercanos que, gracias a la semántica estructural del dominio, serán relevantes para enriquecer la consulta original con conocimiento experto y generar un conjunto de términos completamente adaptados al vocabulario normativo de los documentos oficiales.

Nuestra implementación del *Spreading Activation* consiste en construir dos conjuntos de conceptos que guardan información sobre el estado del algoritmo. Para su definición, tenemos que  $\mathbb{D}_{com}$  es el conjunto de conceptos de la red semántica,  $\Phi^1$  es el conjunto de conceptos inicialmente activados,  $c_j^k$  es un concepto propagado en la iteración  $k$ -ésima (a partir de él se activan otros conceptos) y, por último,  $\omega_{ji}^k$  es la relación por la que se propaga el algoritmo desde  $c_j^k$  hasta  $c_i$  en la iteración  $k$ -ésima:

**Conjunto  $\mathcal{A}$ :** cola de conceptos **activados** candidatos a ser propagados.

Formalmente podemos describirlos de la siguiente manera:

$$\mathcal{A}^0 = \Phi \quad (4.4)$$

$$\mathcal{A}^k = (\mathcal{A}^{k-1} \cup \{c_i : \forall c_j / \omega_{ji}^k > 0\}) - \{\mathcal{G}^k\} \quad (4.5)$$

**Conjunto  $\mathcal{G}$ :** conjunto de conceptos propagados:

$$\mathcal{G}^0 = \emptyset \quad (4.6)$$

$$\mathcal{G}^k = \mathcal{G}^{k-1} \cup \{c_j^k\} \quad (4.7)$$

---

<sup>1</sup> $\Phi$  y  $Q_{sem}$  son el mismo conjunto.

donde el *output* del algoritmo es el conjunto de conceptos ponderados que representa la nueva consulta enriquecida:  $\mathcal{G}^k = Q'_{sem}$ .

Por último, podemos definir el cálculo de la activación de un concepto  $c_i$  en la iteración  $k$ , denotado por  $I_i^k$ . En la iteración 0 podemos calcular el valor de activación de  $c_i$  como:

$$I_i^0 = \begin{cases} 1 & \text{si } c_i \in \Phi \\ 0 & \text{si } c_i \notin \Phi \end{cases} \quad (4.8)$$

en la iteración  $k$ , calculamos el valor de activación  $c_i$  a partir del elemento  $c_j^k$  desde el que se activa.

$$I_i^k = \begin{cases} I_i^{k-1} & \text{si } \omega_{ji}^k = 0 \\ I_i^{k-1} + \omega_{ji}^k I_j^{k-1} & \text{si } \omega_{ji}^k > 0 \end{cases} \quad (4.9)$$

#### 4.4. Nuestra versión de *Spreading Activation*

Estamos trabajando en controlar la activación del algoritmo para obtener el conjunto  $Q'_{sem}$  óptimo para ejecutar la búsqueda sintáctica. A la hora de elegir qué características se debían parametrizar para mejorar el modelo, hemos tenido en cuenta los diferentes factores:

- Esquema conceptual sobre el que se trabaja
- Diseño de las relaciones entre conceptos: ponderación y etiquetado
- Tamaño de la propagación
- Granularidad de los resultados de búsqueda
- Dificultad de implementación
- Evaluación de mejora conseguida como aplicación de las nuevas características

Nuestros esfuerzos actuales se concentran actualmente en los siguientes puntos:

**Degradación de la salida  $O_j$ :** Hemos trabajado en una versión del modelo genérico del *Spreading Activation*, para aplicar algunas restricciones que mejoren su comportamiento y, por tanto, los resultados obtenidos. Actualmente, estamos investigando una redefinición del cálculo del valor de activación a partir de una función  $h$  que degrada el valor de salida  $O_j$  de un concepto  $c_j$ :

- Parametrización genérica: Esta función  $h$  de degradación permite obtener la salida de un concepto  $c_j$  a partir de su grado de activación:

$$O_j = h(I_j) \quad (4.10)$$

- Caso básico: Cuando la función  $h_0$  es igual que la función identidad,  $h_0 = id$ , la salida  $O_j$  toma como valor el nivel de activación del concepto  $c_j$ :

$$O_j = h_0(I_j) = I_j \quad (4.11)$$

- Parametrización usando **distancia**: Como el nivel de activación de un concepto  $n_j$  que ha sido activado desde un concepto inicial  $n_l$  (perteneciente a  $\Phi$ ) depende de la distancia del concepto  $n_l$  al concepto  $n_j$ , cuanto más nos alejemos de  $\Phi$ , la puntuación debe ir decayendo, de forma que para el camino que recorre el algoritmo desde  $n_l$  hasta  $n_j$ :  $I_l > I_j$ .

La función  $h_1$  degrada la salida de los conceptos de forma que disminuye la puntuación de aquellos conceptos que se alejan del “núcleo de activación” y premia cuán cerca esté un concepto de los nodos originales. Así que dada la distancia  $d_j$ , donde  $d_j = \min\{d_{lj} : \forall n_l \in \Phi\}$ :

$$O_j = h_1(I_j, d_j) = \frac{I_j}{d_j} \quad (4.12)$$

- Parametrización usando **pulsos**: Otra forma de degradación del algoritmo es la función  $h_2$ , que se basa, no en la distancia entre los conceptos, sino en la cantidad de iteraciones  $k$  que se han ejecutado:

$$O_j = h_2(I_j, k) = \left(1 + \frac{I_j}{k}\right) \exp\left(-\frac{I_j}{k}\right). \quad (4.13)$$

**Camino:** La red semántica que teje una ontología utiliza relaciones que también están ponderadas. Cada tipo de relación  $\omega_{xy}$  tiene asociado un peso  $w_i$  (p.e. a la relación semántica “is-a” le corresponde un peso de 0,5). Actualmente estamos trabajando en el ajuste de los pesos de relaciones para afinar la relevancia de los conceptos, el grado de activación.

**Nota 4.4.1.** Nuestra hipótesis se basa en que cuanto más específico es un concepto  $c_i$ , es más representativo del dominio y más relevante

*para la búsqueda. Los conceptos más genéricos, como es el caso de categorías de alto nivel, p.e. physical-object, no representan por sí mismos conocimiento específico del dominio. En consecuencia a medida que la propagación asciende en la jerarquía, el nivel de activación de los conceptos desciende.*

El refinamiento y uso de pesos específicos en las relaciones es crítico para la ejecución del algoritmo. El cambio del peso de un relación puede alterar el camino de propagación y los conceptos que antes se activaban ahora no se activan o bien tienen un valor peor de lo esperado.

Finalmente hemos implementado una mejora de **recompensa de caminos**, que se aplica cuando finaliza el algoritmo. La hipótesis es la siguiente: cuando se llega a un concepto o nodo a través de más de un camino y los orígenes de éstos son distintos entonces lo más probable es que ese concepto sea realmente relevante para la consulta del usuario. De ahí que la mejora consista en recompensar a los conceptos que se encuentren en estos caminos, la idea es que si un concepto ha sido propagado por más de una vía debe ser importante por lo que recompensamos a la vía de acceso a ese concepto.

**Definición 4.4.2.** *Sea  $p_i$  el número de caminos que comienzan y terminan en nodos diferentes de  $\Phi^2$  que pasan por el nodo  $c_i$  y sólo contienen nodos pertenecientes a  $\mathcal{G}$ . Esta mejora asigna un nuevo valor de activación a cada nodo  $c_i$  denotado por  $I_i^*$  y se calcula a través de la función  $g$ :*

$$I_i^* = g(I_i, p_i) \quad (4.14)$$

En nuestro caso, tras varias pruebas la definición de la función  $g$  ha sido la siguiente:

$$g(x, y) = x(\log(y + 1) + 1) \quad (4.15)$$

---

<sup>2</sup>La recompensa no se aplica a los nodos que pertenecen a  $\Phi$ .

## Capítulo 5

# Arquitectura del sistema de conocimiento: ontologías y tesauros

En el proyecto BOPA, el diseño de la arquitectura general ha permitido que la información semántica y lingüística, en forma de ontologías y tesauros, haya alcanzado un gran tamaño y que, a la vez, siga siendo manejable. Las consideraciones que se han tenido en cuenta en el diseño de esta nueva arquitectura son las siguientes:

- La conveniencia de poder repartir la tarea de construir, progresivamente y de forma colaborativa entre varias personas, la base de conocimiento.
- La necesidad de dar mantenimiento y ampliar, de forma cómoda, la información.
- El interés en que todos los artefactos construidos puedan ser reutilizados en nuevos proyectos y con diferentes propósitos. La receta para lograrlo es bien conocida: máxima cohesión y mínimo acoplamiento.
- Las limitaciones expresivas de los lenguajes de la Web Semántica para descripción de recursos: RDF(S) y la familia OWL.
- Las limitaciones de las herramientas de edición de ontologías en OWL (Protégé, SWOOP) y de ficheros RDF.
- Las limitaciones de las bibliotecas que permiten procesar las ontologías desde el lenguaje Java (Jena, OWL-API).

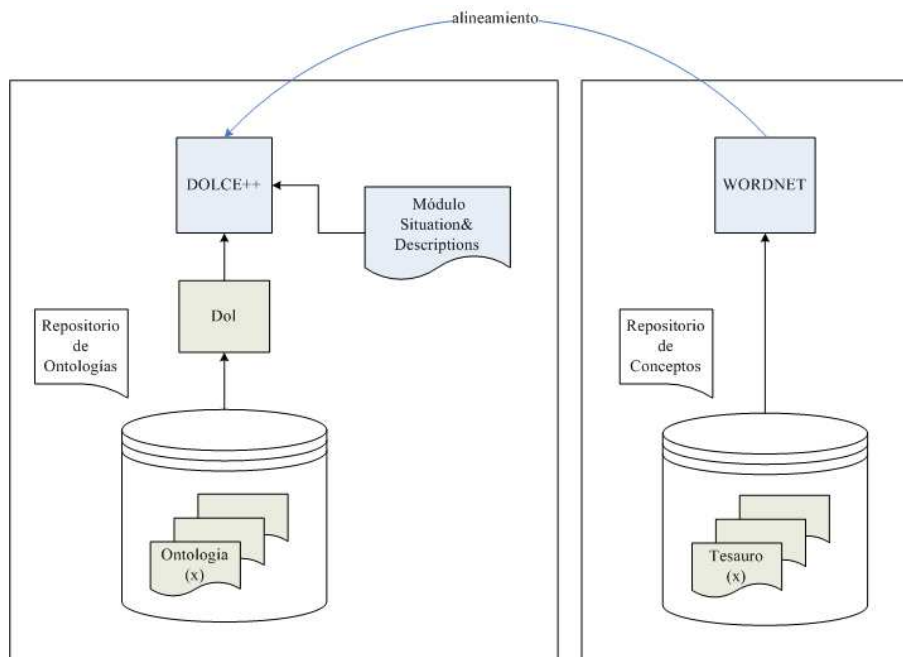


Figura 5.1: Arquitectura física: división en ficheros de la información semántica y léxica.

Por motivos tanto prácticos, como técnicos, como de mantenimiento y de posible reutilización, nos ha parecido necesario una separación física entre el vocabulario y su semántica asociada. Nos interesa, por ejemplo, tratar individualmente el vocabulario o la semántica de un dominio particular. Más aun, esta arquitectura contempla el uso de ficheros de correspondencia intermedios, cuyo objetivo es encapsular la *contextualización* de los términos utilizados en la consulta del usuario. Nuestra arquitectura contempla tres tipos de ficheros:

1. Ontologías.

Hemos utilizado ontologías para la formalización de los dominios. Las ontologías están expresadas en el lenguaje de ontologías propuesto por el W3C, OWL-DL. Como veremos más adelante, se propone una arquitectura modular de las ontologías.

2. Tesauros.

El léxico y el uso de técnicas de *Spreading Activation* ha resultado ser crucial para el éxito del buscador, y, en general, podemos asumir que para cualquier tipo de búsqueda sintáctica.

3. Fichero de correspondencia entre ontologías y tesauros

Las correspondencias entre tesauros y ontologías definen qué concepto del tesoro se identifica con qué clase o instancia en una ontología.

La arquitectura física se representa en el diagrama de la Figura 5.1. En la parte izquierda se encuentran los ficheros de ontologías. En la parte derecha se sitúan los ficheros de tesauros. Entre ambos extremos, el fichero o ficheros de correspondencia relacionan los primeros con los segundos.

Lo que se obtiene con esta arquitectura son distintas bases de conocimiento reutilizables (Figura 5.3): la base de conocimiento léxico (construida con el SKOS y conectada a WordNet) y la base de conocimiento semántico (construida con OWL-DL y alineada con DOLCE).

## 5.1. Tesauros

En nuestro sistema de búsqueda semántica le damos una importancia capital al léxico. La correcta caracterización formal de los dominios es un objetivo evidente del proyecto, pero la gestión del vocabulario es igual o más relevante. El usuario en última instancia va a utilizar palabras como forma de comunicación con el sistema de búsqueda, en nuestro caso “**key-words**”, dado que el interfaz de búsqueda está inspirado en los buscadores sintácticos tradicionales. El reconocimiento y la activación de los conceptos adecuados es crítico para el correcto funcionamiento del sistema, la definición de la función  $\varphi_{\mathbb{D}}$  anterior. Esto implica utilizar técnicas más complejas que simples ficheros de texto con asociaciones entre conceptos de nuestras ontologías y posibles términos de entrada o salida. Tampoco se puede usar la propiedad *rdfs:label* para etiquetar los distintos conceptos ya que no es lo suficientemente flexible para una gestión compleja de un léxico de dimensiones considerables.

Hemos optado por utilizar el SKOS-Core<sup>1</sup>, el vocabulario RDF(S) desarrollado y propuesto del W3C para la construcción de esquemas conceptuales como tesauros, vocabularios controlados, taxonomías de términos, etc.

Una de las ventajas del uso de SKOS es que nos permite tratar como concepto<sup>2</sup> (*skos:concept*) a cualquier clase o instancia de nuestras ontologías y representar la estructura tradicional de términos de un tesoro como un grafo RDF. Más aún, nos permite asociar a cada concepto el conjunto de

---

<sup>1</sup>Las siglas SKOS están por Simple Knowledge Organization System.

<sup>2</sup>Es necesario aclarar aquí el término “concepto”. Un concepto es una construcción mental, una representación de un aspecto del mundo. La confusión en la literatura relacionada con la Web Semántica tiene que ver con algunas tradiciones en Inteligencia Artificial, pero, sobre todo, con el vocabulario de la Lógica Descriptiva. En esta lógica, base de la familia de lenguajes OWL, las clases o predicados monádicos se denominan “**conceptos**”, mientras que las instancias u objetos del dominio son “**individuos**”. Desde un punto de vista estrictamente semántico, y filosófico si se quiere, un concepto es tanto una instancia, como una clase como una propiedad de la aridad que sea.

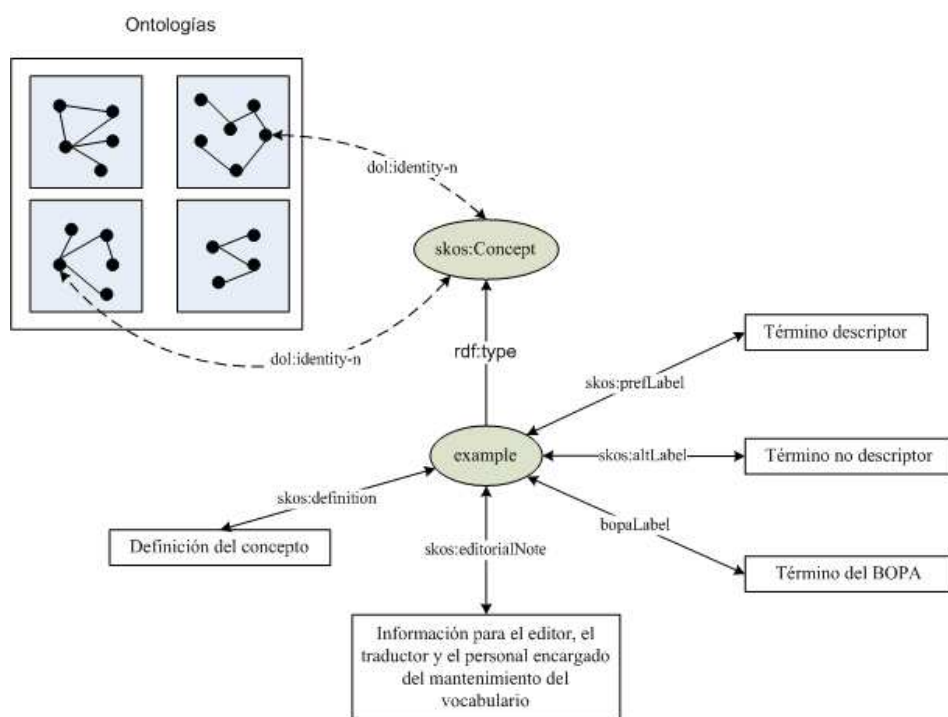


Figura 5.2: Descripción de un concepto utilizando el vocabulario SKOS.

términos sinónimos que se usan para referirse a él, lo que se conoce técnicamente como “synset”.

Nuestro esquema SKOS-Core (véase Figura 5.2) contempla las siguientes relaciones originales del vocabulario:

**Propiedades de etiquetado lingüístico:** las propiedades *skos:prefLabel* y *skos:altLabel* nos permiten construir el *synset* de entrada al sistema. La suma de todos los términos son el input para la ejecución de una búsqueda.

**Propiedades de documentación:** estas propiedades ayudan al mantenimiento del vocabulario, como es el caso de *skos:editorialNote*, que permite al editor del vocabulario controlar la evolución del mismo. Pero también, ayudan a la definición y explicación de los conceptos, tanto a nivel interno como externo: *skos:definition*.

Además de estas propiedades, nuestro tesauro contempla otra propiedad más: *bopa:bopaLabel*. Esta propiedad construye el conjunto de términos de salida para la construcción de la consulta sintáctica  $Q'_{sin}$  a partir del conjunto de conceptos  $Q'_{sem}$ . Este *synset* está completamente adecuado al léxico y vocabulario específico del Boletín.

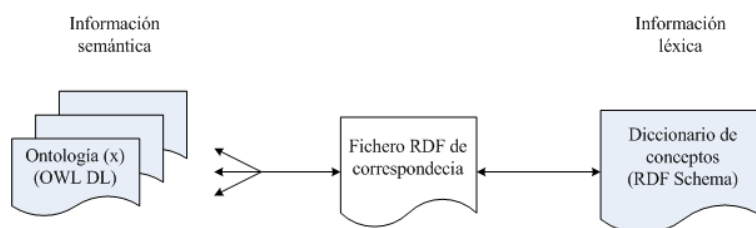


Figura 5.3: Correspondencia entre los ficheros de las ontologías y el tesauro.

De esta forma, a un concepto como puede ser *Persona Mayor*, nuestro SKOS le asocia un conjunto de términos de entrada formado a partir de la información de su *prefLabel* y sus *altLabel*: {“persona mayor”, “abuelo”, “viejo”, “anciano”}. Este conjunto de términos de entrada recoge términos usuales de los ciudadanos, es decir, lenguaje coloquial. Nuestra intención es poder reconocer el mayor número de conceptos en nuestra ontología sin que el ciudadano tenga que modificar su forma de hablar.

Por otro lado, el conjunto de salida de términos de un concepto se construye a partir de la información de sus *bopa:bopaLabel*. Para el caso anterior, está claro que este conjunto de salida está constituido por {“persona mayor”}, el único término de nuestro tesauro que aparece recogido en el vocabulario del Boletín<sup>3</sup>.

Con esta arquitectura, podemos organizar el léxico, en caso de que sea conveniente, de forma independiente de las ontologías mediante la aplicación de técnicas de campos temáticos.

## 5.2. Cuestiones preliminares sobre las ontologías

Independientemente de los problemas tecnológicos a los que hemos tenido que enfrentarnos durante el desarrollo del proyecto, la semántica del buscador está basada en una serie de hipótesis acerca de la naturaleza de las representaciones mentales de los usuarios, el estatus de los conceptos y cómo se organizan.

El tipo de aserciones que hemos asumido como verdaderas tiene un componente metodológico implícito y han servido para guiar el proceso de construcción de las ontologías. La forma en cómo definimos y entendemos tanto a los usuarios como a su forma de percibir el “mundo” de la Administración Pública, se refleja en su diseño.

<sup>3</sup>Como se aprecia en el ejemplo, el conjunto de términos de salida debe ser un subconjunto del conjunto de términos de entrada.

**Definición 5.2.1.** *Cada ontología de dominio  $O_i$  es la organización relevante de los conceptos que forman los posibles conjuntos de preguntas de un usuario sobre el dominio  $\mathbb{D}_i$ .*

Hemos establecido internamente una serie de principios estructurales. En la medida en que nuestras ontologías se adecúen lo más posible a estos principios, más cerca estaremos del razonamiento intuitivo del usuario.

- **Principios sobre los conceptos**

**Nota 5.2.2.** *Los conceptos son representaciones mentales que permiten describir y clasificar determinados objetos y fenómenos del mundo.*

**Nota 5.2.3.** *Las situaciones determinan qué datos son los más fácilmente accesibles por los sujetos, y por tanto, cuáles son los conceptos recuperables en cada momento. Analizar las posibles situaciones de búsqueda de los usuarios y qué objetivos les mueven, nos ayuda a construir el conjunto de conceptos que deben ser estructurados mediante una ontología.*

**Nota 5.2.4.** *Los conceptos pueden ser representados formalmente por clases o instancias en una ontología.*

- **Principios sobre las ontologías**

**Nota 5.2.5.** *Una ontología de dominio se acerca más a nuestra manera de actuar y está estructurada de una manera más próxima al usuario. Nuestra mente funciona mejor con representaciones localmente eficaces que con representaciones inespecíficas.*

**Nota 5.2.6.** *Muchos niveles de jerarquía dan lugar a herencias demasiado ricas que pueden introducir en las clases información redundante, en ocasiones irrelevante o contradictoria.*

**Nota 5.2.7.** *El uso de las técnicas de Spreading Activation en la búsqueda de conceptos semánticamente próximos tenderá a ser más efectiva cuanto más restringidos y específicos sean los contextos de búsqueda.*

### 5.3. *Framework* semántico

Para poder integrar, evaluar y, en definitiva, trabajar con varias ontologías a la vez, es necesario un marco semántico que establezca una estructura común básica. Nuestra arquitectura sostiene un sistema constituido por varias ontologías. Su esqueleto está basado en una ontología fundacional<sup>4</sup>

---

<sup>4</sup>Este tipo de ontologías también se conocen como ontologías de alto nivel.

que organiza de forma homogénea la semántica de los dominios y los formaliza de acuerdo a unos conceptos generales compartidos por todas las ontologías del sistema.

Utilizar un marco semántico de este tipo asegura dos cosas: una línea metodológica y una semántica interoperable. Definitivamente, hemos optado por la ontología de alto nivel DOLCE (the Descriptive Ontology for Linguistic and Cognitive Engineering) desarrollada y mantenida en el LOA (Laboratory for Applied Ontology). DOLCE tiene una clara tendencia cognitiva. Su intención es capturar las categorías ontológicas que subyacen al lenguaje natural y al sentido común humano. Lo interesante de DOLCE es que no adopta una postura **prescriptiva** y no obliga a una visión unilateral sobre la naturaleza intrínseca del mundo. Las categorías de DOLCE, al contrario, son artefactos cognitivos, que dependen en última instancia de la percepción humana, las improntas culturales y las convecciones sociales.

El dominio de cuantificación de DOLCE está constituido por las **entidades particulares**. Los **universales** o predicados no forman parte del universo de discurso, aunque se usan para organizar y caracterizar los particulares. En otras palabras, DOLCE es una teoría formal de primer orden sobre cuáles son las categorías primitivas de la percepción y cognición de la mente humana. Como en la mayoría de las ontologías formales, se asumen las siguientes categorías que dividen y organizan el mundo de los particulares:

**Endurant** Los *endurant* son particulares en el espacio, que participan como mínimo en un *perdurant*. Los *endurant* pueden ser físicos (p.e. una piedra) o no físicos (p.e. una compañía de seguros), y agentivos (p.e. el portero de mi casa) o no agentivos (p.e. un libro de matemáticas).

**Perdurant** Los *perdurant* son particulares en el tiempo, que tienen como mínimo un participante. Son entidades que existen parcialmente en un momento determinado del tiempo (p.e. estados, eventos, procesos, etc.). Su existencia e identidad completa se extiende a lo largo de un intervalo de tiempo  $t_i$ .

**Quality** Los *quality* son particulares dependientes de otras entidades, cualidades “inherentes” tanto en *endurants* como *perdurants* (p.e. los colores, las velocidades, etc.)

Hemos utilizado también el módulo de las “Descriptions and Situations” (DnS), desarrollado por Aldo Gangemi (fix). Esta ontología es un mecanismo muy potente para extender el alcance de DOLCE, y en general de cualquier ontología. DnS permite la reificación de cualquier entidad imaginable, desde predicados, sistemas, dominios enteros y, en general, cualquier tipo de organización compleja del mundo. El propósito fundamental

de DnS es permitir una caracterización asequible de los objetos sociales y muy especialmente, los objetos informativos, las organizaciones y lo que se conoce como “descripciones” o conceptualizaciones (es el caso de normas, planes, métodos, diseños de artefactos, etc.). El módulo DnS es especialmente necesario en nuestras ontologías al tener que enfrentarnos con el modelado de situaciones reguladas por normas como reglamentos, leyes o regulación administrativa, y con el modelado de la Administración Pública, que, en última instancia, es un sistema articulado por consejerías, servicios, competencias territoriales, colectivos de funcionarios, procedimientos, edificios y lugares físicos, etc.

DnS proporciona un vocabulario de símbolos de predicado y un conjunto de axiomas para tipificar los nuevos individuos, interrelacionarlos y conectarlos con el resto de los predicados DOLCE. El módulo DnS aumenta su vocabulario y genera lo que se conoce como DOLCE+. En nuestro buscador, hemos utilizado una versión reducida de DOLCE+ construida *ad hoc*. Hemos considerado qué conceptos de DOLCE+ son interesantes para el modelado de nuestros dominios y lo hemos extendido en algunos puntos utilizando vocabularios como el Dublin Core o desarrollando categorías como los *roles*, los *objetos informativos* y las *organizaciones*. El objetivo es convertir esta ontología, a la que hemos llamado *Dol*, en un marco semántico.

Para explotar *Dol* como si fuese un sistema de tipos, de forma que nos garantice la consistencia de cada ontología  $O_i$  de dominio, hemos establecido la siguiente metodología.

**Definición 5.3.1.** *Sea  $O$  una teoría axiomática de primer orden tipificada sobre un dominio  $\mathbb{D}_O$ , el vocabulario  $\Sigma_O$  de  $O$  está formado por la tupla  $\Sigma_O = \langle \Delta_O, \Phi_O, \Omega_O \rangle$ , donde  $\Delta_O$ ,  $\Phi_O$  y  $\Omega_O$  son conjuntos disjuntos:*

- $\Delta_O$  el conjunto de clases de  $O$ .
- $\Phi_O$  el conjunto de individuos de  $O$ .
- $\Omega_O$  el conjunto de predicados binarios de  $O$ .

Los conceptos  $c_i$  de  $\mathbb{D}_O$  constituyen parte del vocabulario de  $O$  ( $\mathbb{D}_O \in \Sigma_O$ ). De forma que para el dominio  $\mathbb{D}_O$  y la correspondiente ontología  $O$ , se cumple que:

$$\mathbb{D}_O = \Delta_O \cup \Phi_O \quad (5.1)$$

$$\text{si } c_i \in \mathbb{D}_O, \text{ entonces } \begin{cases} c_i \in \Delta_O & \text{si } c_i \text{ es un universal} \\ c_i \in \Phi_O & \text{si } c_i \text{ es un particular} \end{cases} \quad (5.2)$$

Además, para que *Dol* se convierta en un verdadero marco de desarrollo tenemos que poder construir el modelo de cada dominio a partir de *Dol*.

**Definición 5.3.2.** *En esta arquitectura, dada la ontología fundacional  $Dol$ , la ontología de dominio  $O$  y el operador  $owl:imports$  del lenguaje OWL ,*

$$\left\{ \begin{array}{l} O \text{ owl:imports } Dol \Rightarrow O_{dol} \\ \Sigma_{O_{dol}} = (\Sigma_O \cup \Sigma_{Dol}) \end{array} \right. \quad (5.3)$$

Si ahora utilizamos la teoría de modelos para proporcionar una interpretación semántica  $\mathfrak{A} = \langle \mathcal{W}, \mathcal{F} \rangle$  a la ontología  $O_{dol}$ , entonces  $\mathcal{W}$  es el universo de interpretación y  $\mathcal{F}$  la función de interpretación, cuyo dominio es el vocabulario de  $O_{dol}$ . Dado que el vocabulario  $\Sigma_{O_{dol}} = (\Sigma_O \cup \Sigma_{Dol})$ , podemos afirmar que  $\mathcal{W}_O$  es el dominio donde se interpretan los símbolos de  $\Sigma_O$ , donde  $\mathcal{W}_O \subset \mathcal{W}$ , y que  $\mathcal{W}_{Dol}$  es el dominio donde se interpretan los símbolos de  $\Sigma_{Dol}$ , donde  $\mathcal{W}_{Dol} \subset \mathcal{W}$ .

**Definición 5.3.3.** *Así que la función de interpretación  $\mathcal{F}$  para cada conjunto de símbolos del vocabulario  $\Sigma_{O_{dol}}$  se define de la forma siguiente:*

$$\mathcal{F} : \left\{ \begin{array}{l} \langle \Delta_O, \Phi_O, \Omega_O \rangle \longrightarrow \langle \mathcal{W}_O, \mathcal{W}_O, \mathcal{W}_O \times \mathcal{W}_O \rangle \\ \langle \Delta_{Dol}, \Phi_{Dol}, \Omega_{Dol} \rangle \longrightarrow \langle \mathcal{W}_{Dol}, \mathcal{W}_{Dol}, \mathcal{W}_{Dol} \times \mathcal{W}_{Dol} \rangle \end{array} \right. \quad (5.4)$$

Hemos establecido además una serie de restricciones en la operación  $owl:imports$  para garantizar la consistencia de nuestra metodología. Nos interesa que la formación de cada ontología de dominio  $O_{dol}$  sea consistente y homogénea dentro del sistema, y que  $O_{dol}$  sea una extensión monotónica de  $Dol$ .

**Restricción 1:**  $\mathcal{W}_O \subset \mathcal{W}_{Dol}$

**Restricción 2:**  $\Omega_{O_{dol}} = \Omega_{Dol} (\Rightarrow \Omega_O = \emptyset)$

**Restricción 3:**  $\Phi_{O_{dol}} = \Phi_O (\Rightarrow \Phi_{Dol} = \emptyset)$

Estas restricciones determinan una serie de características de las ontologías de dominio:

**Nota 5.3.4.** *Todos los individuos pertenecen a  $O$ , son objetos del dominio.*

**Nota 5.3.5.** *Todas las relaciones entre individuos pertenecen a  $Dol$ . Esta ontología nos permite clasificar y organizar a los particulares de cada dominio mediante las mismas relaciones (basadas en teorías formales y filosóficas sobre la estructura del mundo).*

**Nota 5.3.6.** *Un dominio  $O$  es una extensión monotónica de  $Dol$ , de forma que*

$$\forall A \in \Delta_O \Rightarrow \begin{cases} \exists B \in \Delta_{Dol} / A \sqsubseteq B \\ \vee \\ \exists C \in \Delta_O, \exists B \in \Delta_{Dol} / A \sqsubseteq \dots \sqsubseteq C \sqsubseteq B \end{cases} \quad (5.5)$$

## 5.4. Metodología y conceptos generales

Hemos utilizado la metodología OntoClean para construir las ontologías de **dominio**. OntoClean nos ha servido para validar la adecuación de las relaciones *is-a*. Ontoclean es un conjunto de metapropiedades para guiar y evaluar el desarrollo de construcción de ontologías basadas en las categorías básicas de análisis de la filosofía: la *identidad*, *dependencia*, *rigidez* y *unidad*. Estas metapropiedades imponen ciertas restricciones sobre la relación de subsunción que ayuda a clarificar y a ordenar coherentemente las taxonomías de conceptos.

*Rigidez*(**R**): Una propiedad es rígida si es esencial para todas sus posibles instancias; una instancia de una propiedad rígida no puede dejar de ser instancia de esa propiedad en un mundo diferente. Una propiedad en una ontología puede ser rígida(+**R**), no rígida (−**R**) o antirrígida (∼**R**).

*Unidad*(**U**): La unidad se refiere al problema de describir las partes y los límites de los objetos, de forma que sabemos en general qué es parte de un objeto, qué no es, y bajo qué condiciones el objeto es un *todo*. Las condiciones específicas que determinan cómo se relacionan las partes y el todo, se conocen como *criterio de unidad*. Las propiedades cuyas instancias sean *todos* pueden tener criterios de unidad distintos. En general, una propiedad cualquiera de una ontología puede ser unitaria (+**R**), no unitaria (−**R**) o antiunitaria (∼**R**).

*Identidad* (**I**): El problema filosófico de la identidad se refiere al problema de reconocer entidades individuales como el mismo individuo (o diferente). En OntoClean se distinguen aquellas propiedades que arrastran un *criterio de identidad*<sup>5</sup>. También se diferencia en OntoClean entre las entidades (+**I**) y aquellas que no (−**I**). El problema de la identidad es bastante complicado, ya que implica identificar y detectar las

---

<sup>5</sup>Debido a que el criterio de identidad se hereda a través de la jerarquía de propiedades, OntoClean ha añadido una distinción de más alcance para aquellas propiedades que proporcionan un criterio de identidad “propio”( +**O**), no heredado por la subsunción de propiedades.

propiedades *esenciales* de un individuo. Obviamente, dos cosas que no compartan las mismas propiedades esenciales no pueden ser idénticas.

*Dependencia (D)*: La dependencia ontológica es una relación entre entidades. Hay varios tipos de dependencias. En nuestro caso, vamos a considerar la dependencia de un individuo en otras entidades para su existencia e identidad. La metapropiedad de dependencia se define formalmente a partir de la relación de *constitución*. Por ejemplo, una persona y el cuerpo humano que lo constituye tienen las mismas partes, pero una persona no puede existir sin un cuerpo humano. En este sentido, depende de él.

A continuación, señalamos cuáles son las restricciones que se siguen de las definiciones de las cuatro metapropiedades anteriores. Para una introducción y un desarrollo más extenso de esta presentación (FIX). Sean  $\gamma$  y  $\varphi$  dos propiedades cualesquiera:

$$\gamma^{\sim R} \text{ no puede subsumir } \varphi^{+R} \quad (5.6)$$

$$\gamma^{+I} \text{ no puede subsumir } \varphi^{-I} \quad (5.7)$$

$$\gamma^{+U} \text{ no puede subsumir } \varphi^{-U} \quad (5.8)$$

$$\gamma^{\sim U} \text{ no puede subsumir } \varphi^{+U} \quad (5.9)$$

$$\gamma^{+D} \text{ no puede subsumir } \varphi^{-D} \quad (5.10)$$

La construcción del modelo formal de los dominios que cubre el BOPA se ha basado tanto en el uso de *Dol* como en la metodología OntoClean, consistente con el propio uso de DOLCE. Además, siendo las ontologías una especificación de una conceptualización, el **significado** que connotan los términos en las ontologías de dominio está ligado al uso interno que hace de ellos una comunidad determinada y, por tanto, al igual que ocurre con las ontologías fundacionales, debe ser el resultado de un proceso intersubjetivo (FENSEL 2004 FIX). Para alcanzar este objetivo, y más tratándose de información oficial del Principado de Asturias, hemos tomado como punto de partida la legislación vigente en España y, más concretamente, en la comunidad autónoma del Principado de Asturias. Hemos tenido en cuenta, además, la legislación europea y hemos utilizado como principal fuente de datos léxicos los sistemas de clasificación<sup>6</sup> procedentes tanto del Instituto Nacional de Estadística como del EUROSTAT, en particular:

- CNO-INEM: el sistema de Clasificación Nacional de Ocupaciones Profesionales (CNO) extendido por el Servicio de Empleo Estatal (INEM).

---

<sup>6</sup>Estas clasificaciones son fundamentales en el modelado de conocimiento de la Administración Pública ya en su mayor caso, aparte de ser utilizadas por los Institutos Nacionales de Estadística, su uso es normativo en la codificación de los procedimientos normativos, como es el caso de la contratación pública.

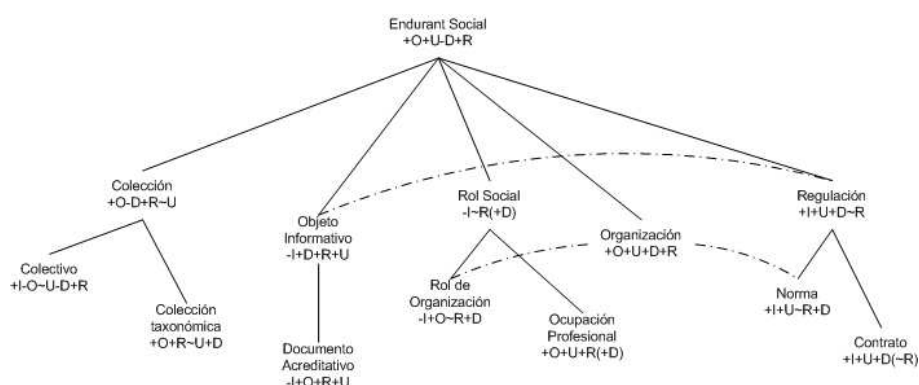


Figura 5.4: FIX.

- CPV: *Common Procurement Vocabulary* es la clasificación usada obligatoriamente a nivel europeo para la clasificación de los objetos de los contratos públicos y cualquier fase del procedimiento de licitación en general.

Entre los conceptos que son piedra angular de nuestras ontologías presentamos a continuación los más representativos y sobre los que descansa la organización semántica de cada ontología. La metodología OntoClean está aplicada en la Figura 5.4:

*Colección*: Una colección es un objeto social que aglutina cualquier tipo de entidad (eventos, objetos, personas, etc.). En realidad, una colección no es más que la reificación de un conjunto de objetos definido por extensión (en otras palabras, la reificación de la extensión de una clase). Por ejemplo, si en mi casa tengo cuatro sellos de la época franquista, cada uno de ellos pertenece a la clase *sello*, y además esos cuatro sellos forman una entidad por sí mismos, esa entidad es una colección, en particular, mi colección de sellos. Utilizamos las colecciones, sobre todo, para interpretar las categorías de las clasificaciones internacionales, como es el caso de la CPV. También el Boletín Oficial del Principado de Asturias es una colección, una colección de textos cuyos miembros son todas y cada de las disposiciones publicadas.

*Colectivo*: un colectivo es un tipo especial colección en la sólo los agentes pueden ser miembros. Dentro de nuestro inventario, los colectivos organizan las categorías de las clasificaciones de empresas y las de las ocupaciones profesionales (CNO-INEM). También son colectivos los cuerpos y escalas de los funcionarios del Principado de Asturias, y grupos sociales como la familia o las minorías sociales.

*Rol*: un rol es una entidad social que describe propiedades no esenciales de particulares. Los roles permiten clasificarlos y usarlos en descripciones de estados y hechos del mundo. En nuestro modelo de la Administración Pública, el destinatario de una subvención, el objeto de un contrato público y los distintos puestos que puede desempeñar un empleado público, como jefe de servicio o recadero, son roles. Así hemos desarrollado una clasificación de los distintos roles que puede “jugar” un miembro de una organización. También hemos tratado la ocupación profesional como un rol. La CNO-INEM adquiere de esta manera un modelado más particular que el resto de las clasificaciones.

*Organización*: Una organización es una persona construida socialmente y con una articulación interna compleja: tareas, roles dentro de la organización y otras figuras. Una organización tiene un estatus legal y una serie de poderes que le confieren la ley y sus estatutos. Las organizaciones son compañías, departamentos, asociaciones, la universidad y cómo no, los órganos de la Administración Pública. El modelado de la estructura de la organización es el pilar de nuestras ontologías: qué papel juegan los empleados del Principado dentro de su estructura, cuál es la clasificación de los organismos administrativos territoriales, qué es una consejería, un servicio o entidades locales como las mancomunidades, o cuál es la relación entre los territorios sobre los que una organización tiene competencia o jurisdicción, y cómo se interpretan los edificios y demás complejos urbanísticos como sede social de organizaciones.

*Objeto informativo*: los objetos informativos son entidades sociales y abstractas, que se materializan en algún objeto: papel, formato digital, etc. Los objetos informativos expresan y codifican una determinada información y es el tipo de información que codifican lo que permite distinguirlos y categorizarlos. Son objetos informativos las disposiciones del BOPA, pero también lo son los documentos acreditativos como el DNI, un pasaporte o la titulación académica. Ésta última es de vital importancia para el dominio del empleo público, donde las ocupaciones profesionales requieren teóricamente una determinada cualificación laboral.

## 5.5. Organización de las ontologías

Las ontologías se organizan dentro de un sistema complejo de conocimiento de distintos niveles. Cuando hablamos de “arquitectura semántica”, la expresión puede tener varias interpretaciones: la estructura modular que utilizamos para segmentar los ámbitos de actuación de la Administración

Pública, las posibles divisiones dentro de cada uno de estos dominios o la propia arquitectura global en la que se utiliza el marco *Dol*. Intentaremos dar una visión general del asunto que nos concierne.

**Nota 5.5.1.** *Nuestro sistema tiene que inferir los objetivos del usuario expresados en la consulta ( $Q_{sin}$ ) y traducirlos en una expresión ( $Q_{sem}$ ) compatible con la semántica de nuestras ontologías.*

El conocimiento está dividido en dos bloques cualitativamente distintos:

1. La ontología del Boletín Oficial del Principado de Asturias, que formaliza la estructura básica del Boletín y de los documentos que la conforman. Nos permite clasificar y describir correctamente los documentos organizando sus metadatos en un sistema facetado, y además realizar inferencias básicas para filtrar el conjunto total de resultados obtenidos para una consulta. La ontología estructura y relaciona las categorías básicas de la Administración Pública, los organismos territoriales administrativos (como son las consejerías y los ayuntamientos), con las propiedades básicas los documentos del Boletín. En el nivel extensional, la ontología organiza los distintos tipos de disposición oficial<sup>7</sup>.

Como ya comentamos en la sección 2, la relación intensional entre un documento y el contenido que expresa se encauza a través de la relación entre el documento y las ontologías que se utilizan para buscar sobre él. Al carecer de etiquetado con los conceptos del dominio, esta relación es dinámica, según se vayan ejecutando las consultas del usuario, y sintáctica, se busca el “matching” entre las palabras del documento y el tesoro de la ontología de dominio.

Esta parte, la representación de las competencias, es la más complicada. Cada ontología de dominio es una definición compleja de una competencia administrativa, reificada<sup>8</sup> dentro de la ontología del Boletín Oficial del Principado de Asturias.

2. Las ontologías de dominio, que formalizan las competencias de la administración y nos proporcionan las semánticas sobre las que ejecutar las técnicas de *Spreading Activation*.

---

<sup>7</sup>Desgraciadamente, los documentos vienen etiquetados genéricamente sin tener en cuenta su auténtico significado: “Anuncios” o “Resoluciones”. Para extraer la información correspondiente a cada artículo, por ejemplo, si es un concurso o un reglamento, hemos desarrollado una herramienta de clasificación semi-automática de textos.

<sup>8</sup>La reificación de una ontología, como entidad de primer orden, dentro de otra ontología es un problema de tercer orden, fuera del alcance de este documento. Aunque la teoría DnS proporcione funciones de reificación, necesitamos un sistema lógico de transformación más expresivo para poder explicar la reificación de las propias ontologías como entidades de primer orden en otras ontologías que las contienen.

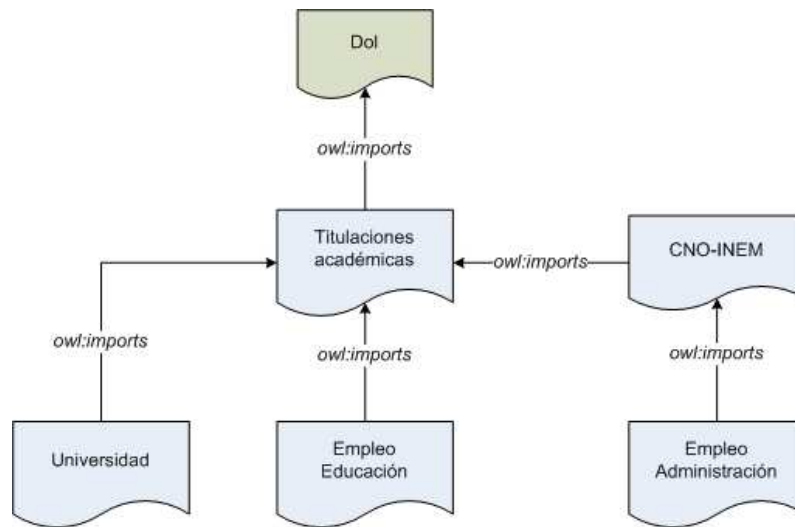


Figura 5.5: Dominio del empleo público

El ámbito administrativo es muy extenso. Está además sujeto a distintas interpretaciones, como es el caso por ejemplo de las leyes, que regulan determinados aspectos del “mundo” contemplando solamente determinadas propiedades y relaciones de las entidades. Por eso, hemos adoptado una perspectiva modular para la organización del conocimiento sobre los distintos dominios. Pero incluso compartimentar los distintos posibles ámbitos de actuación de las Administración puede no ser suficiente. Depende del nivel de detalle en el que trabajemos, podemos estar ante ontologías con cientos de conceptos y miles de instancias. Por eso, podemos encontrarnos con dominios, como veremos inmediatamente, en los que fue necesario construir varias ontologías complementarias para modelar de una forma flexible la información disponible. Esta división nos permite añadir nuevas ontologías de distinto grano, reutilizarlas, focalizar la representación en distintos puntos del dominio, darles mantenimiento de forma sencilla y controlar más fácilmente la consistencia de la ontología<sup>9</sup>. Entre otras razones

1. Se obtiene un sistema escalable. Podemos ir enchufando sucesivamente nuevos dominios sin romper el diseño y consistencia del sistema.
2. Se puede obtener un mejor rendimiento en tiempo de ejecución si se puede aislar la parte del conocimiento con la que se desea trabajar. Esto permite aplicar políticas de jerarquía de memoria.

<sup>9</sup>Estas razones son las mismas por las que habitualmente se divide un fichero de código fuente demasiado grande en otros más pequeños.

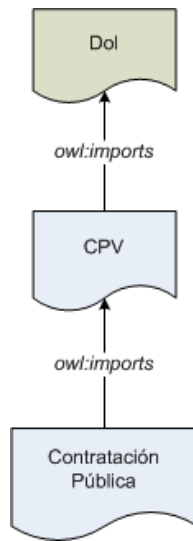


Figura 5.6: Dominio de la contratación pública

3. Nuestra filosofía implica restringir la activación de los conceptos. Al estar las ontologías destinadas a ser consumidas por la aplicación con unos fines específicos (en particular, la activación de conceptos para generar consultas) es importante controlar el tipo de contextos que se activan para evitar salirse de objetivos de búsqueda del usuario. Por ejemplo, si una persona está buscando una plaza de oposición para educación secundaria, difícilmente estará interesado en una plaza de bombero, aunque ambos sean personal de la administración y aparezcan dentro del dominio del empleo público, y por supuesto, menos interesado estará aún en las subvenciones para los ganaderos.

Hasta ahora hemos trabajado sobre tres dominios de la Administración: la contratación pública, las subvenciones y ayudas y el empleo público. A continuación presentamos un breve resumen:

**Dominio de empleo público (véase Figura 5.5)** Constituye la relación entre el sistema de clasificación de ocupaciones profesionales usado por el Servicio de Empleo Estatal (INEM), la Clasificación Nacional de Ocupaciones (CNO) tal y como la extiende el propio INEM (clasificación CNO/INEM), las titulaciones requeridas para cada titulación y los distintos puestos de dentro de la Administración Pública: el colectivo de funcionarios que están relacionados con los diferentes cuerpos y escalas del personal funcionario y el Personal de Administración y Servicios de la Administración Pública (PAS). Este dominio también incluye los puestos de trabajo relacionados con la educación, tanto la enseñanza media como la enseñanza superior, la universidad y por

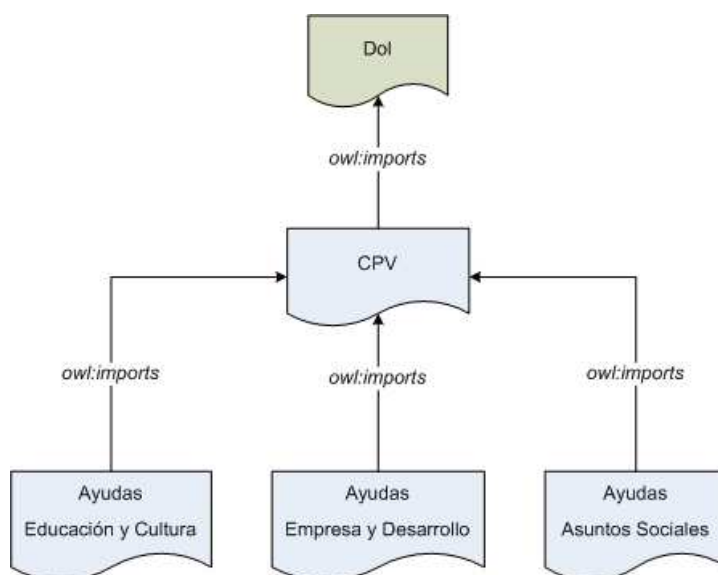


Figura 5.7: Dominio de las ayudas públicas

supuesto los distintos cuerpos de funcionarios docentes.

**Dominio de la contratación pública (véase Figura 5.6)** Este dominio gira en torno a la definición del contrato público. La formalización recoge la identificación y modelado de los roles que desempeñan los participantes en un contrato público (contratante y contratista), el procedimiento de licitación de un contrato y los actos administrativos en los que se subdivide, los tipos de contrato, las personas jurídicas que participan en el procedimiento y las características del propio procedimiento, tanto en su forma como en su tramitación.

Hemos utilizado la clasificación oficial de empresas que recoge el *Texto Refundido de la Ley de Contratos de las Administraciones Públicas* para el contratante, y las Clasificación de Objetos de Contrato (CPV), para unificar las referencias de los objetos de contrato tanto para los órganos de contratación como para las entidades adjudicadoras.

**Dominio de las ayudas públicas (véase Figura 5.7)** En este dominio tratamos a las ayudas públicas como los movimientos de dinero o exenciones de impuestos que, gestionados por la Administración Pública, se entregan a favor de particulares u otras entidades públicas o privadas para realizar alguna actividad o acometer un proyecto específico sin que exista una contrapartida por parte de los beneficiarios.

La dificultad práctica que presenta la división del dominio se debe principalmente a la gran diversidad de objetos subvencionables, así como

a la cantidad de posibles destinatarios. Esto como en el caso del empleo público, incitó la construcción de varias ontologías que modelan diferentes tipos de ayudas

# Bibliografía

- [1] M. Agosti and F. Crestani. A methodology for the automatic construction of a hypertext for information retrieval. In *SAC '93: Proceedings of the 1993 ACM/SIGAPP symposium on Applied computing*, pages 745–753, New York, NY, USA, 1993. ACM Press.
- [2] M. Annamalai and L. Sterling. Guidelines for constructing reusable domain ontologies. In *Ontologies in Agent Systems 2003, Proceedings of the Workshop on Ontologies in Agent Systems (OAS 2003) at the 2nd International Joint Conference on Autonomous Agents and Multi-Agent Systems, Melbourne, Australia, July 15, 2003*, pages 71–74, 2003.
- [3] G. Antoniou and F. van Harmelen. *A Semantic Web Primer*. MIT Press, Cambridge, 2004.
- [4] F. Baader, I. Horrocks, and U. Sattler. Description logics as ontology languages for the semantic web. In D. Hutter and W. Stephan, editors, *Mechanizing Mathematical Reasoning, Essays in Honor of Jorg H. Siekmann on the Occasion of His 60th Birthday*, volume 2605 of *Lecture Notes in Computer Science*, pages 228–248, Berlin, 2005. Springer.
- [5] R. A. Baeza-Yates and B. A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.
- [6] S. Bechhofer, L. Carr, C. A. Goble, S. Kampa, and T. Miles-Board. The semantics of semantic annotation. In *On the Move to Meaningful Internet Systems*, volume 2519 of *Lecture Notes in Computer Science*, pages 1152–1167. Springer, 2002.
- [7] S. Bechhofer and I. Horrocks. The wonderweb ontology. language layer. report and tutorial. Deliverable D1, WonderWeb. Ontology Infrastructure for the Semantic Web, 2006.02.07 2003.
- [8] S. Bechhofer, R. Volz, and P. W. Lord. Cooking the Semantic Web with the OWL API. In *International Semantic Web Conference*, pages 659–675, 2003.

- [9] K. Beck. *Extreme Programming Explained: Embrace Change*. Addison-Wesley Professional, 1999.
- [10] B. Berendt, A. Hotho, and G. Stumme. Towards semantic web mining. In I. Horrocks and J. Hendler, editors, *International Semantic Web Conference (ISWC 2002)*, 2002.
- [11] H. Berger, M. Dittenbach, and D. Merkl. An adaptive information retrieval system based on associative networks. In *CRPIT '04: Proceedings of the first Asian-Pacific conference on Conceptual modelling*, pages 27–36, Darlinghurst, Australia, Australia, 2004. Australian Computer Society, Inc.
- [12] G. Boella, L. Lesmo, and R. Damiano. On the ontological status of norms. In V. R. Benjamins, P. Casanovas, J. Breuker, and A. Gangemi, editors, *Law and the Semantic Web: Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications*, pages 125–141, 2003.
- [13] S. Borgo and C. Masolo. Qualities in possible worlds. In *Proceedings of the Fourth International Conference FOIS 2006*, Baltimore, Maryland (USA), 2006.
- [14] E. Bottazzi and R. Ferrario. Preliminaries of a dolce ontology of organizations. *International Journal of Business Process Integration and Management*, 1?(1/2/3?):64–74?, 2006?
- [15] J. Breuker, A. Elhag, E. Petkov, and R. Winkels. Ontologies for legal information serving and knowledge management. In T. Bench-Capon, A. Daskalopulu, and R. Winkels, editors, *Legal Knowledge and Information Systems*. IOS Press, 2002.
- [16] P. Castells, E. Pulido, C. Carranza, M. Rico, F. Perdrix, E. Piqué, J. Cal, R. Benjamins, J. Contreras, J. Lorés, and T. Granollers. Neptuno: tecnologías de la web semántica para una hemeroteca digital. In R. Navarro-Prieto and J. Lorés-Vidal, editors, *HCI related papers of Interacción 2004*, Lleida, Mayo 2004.
- [17] H. Chen and T. Ng. An algorithmic approach to concept exploration in a large knowledge network (automatic thesaurus consultation): symbolic branch-and-bound search vs. connectionist hopfield net activation. *J. Am. Soc. Inf. Sci.*, 46(5):348–369, 1995.
- [18] P. R. Cohen and R. Kjeldsen. Information retrieval by constrained spreading activation in semantic networks. *Inf. Process. Manage.*, 23(4):255–268, 1987.

- [19] S. Comai, E. Damiani, and L. Tanca. Semantics-aware querying in the WWW: The WG-Log web query system. In *ICMCS, Vol. 2*, pages 317–322, 1999.
- [20] F. Crestani. Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, (11):453–482, 1997.
- [21] H. Cui, J. Wen, J.Ñie, and W. Ma. Query expansion by mining user logs. *IEEE Transaction on Knowledge and Data Engineering*, 15(4):829–839, July 2003.
- [22] D. M. D. Alur, J. Crupi. *Core J2EE Patterns: Best Practices and Design Strategies*. Sun Microsystems, 2003.
- [23] I. de Empleo. Servicio Público de Empleo Estatal. *Clasificación de Ocupaciones*. Servicio Público de Empleo Estatal. Subdirección General de Promoción de Empleo., España, 2004.
- [24] S. Despres and S. Szulman. Construction of a legal ontology from a european community legislative text. In T. Gordon, editor, *Legal Knowledge and Information Systems*. IOS Press, 2004.
- [25] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.
- [26] D. Fensel. *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce*. Springer, Berlin, 2004.
- [27] A. Gangemi, N. Guarino, C. Masolo, and A. Oltramari. Understanding top-level ontological distinctions. In A. G. Pérez, M. Gruninger, H. Stuckenschmidt, and M. Ushold, editors, *Proceedings of IJCAI 2001 Workshop on Ontologies and Information Sharing*, Seattle, Washington, August 4-5 2001.
- [28] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider. Sweetening ontologies with dolce. In *EKAW*, 2002.
- [29] A. Gangemi and P. Mika. Understanding the semantic web through descriptions and situations. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE - OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003*, pages 689–706, 2003.
- [30] A. Gangemi, R.Ñavigli, and P. Velardi. Axiomatizing wordnet glosses in the ontowordnet project. In *Workshop on Human Language Technology for the Semantic Web and Web Services, 2nd International Semantic Web Conference ( ISWC2003)*, Sanibel Island, Florida, October 20-23 2003.

- [31] A. Gangemi, M.-T. Sagri, and D. Tiscornia. Metadata for content description in legal information. In *Workshop on Legal Ontologies, 9th International Conference on Artificial Intelligence and Law (ICAIL-2003)*, Edinburgh, UK, June 24-28 2003.
- [32] S. Ghita, W.Ñejdl, and R. Paiu. Semantically rich recommendations in social networks for sharing and exchanging semantic context. In *Proceedings of the 2nd European Semantic Web Conference Workshop on Ontologies in P2P Communities*, 2005.
- [33] B. Gil-Urdiciain. *Manual de lenguajes documentales*. Ediciones Trea, Gijón, Asturias, 2004.
- [34] A. Gómez-Pérez, M. Fernández-López, and O. Corcho. *Ontological Engineering with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer, Berlin, 2004.
- [35] O. Gospodnetić and E. Hatcher. *Lucene in Action*. Manning, 2005.
- [36] B. C. Grau, B. Parsia, and E. Sirin. Working with multiple ontologies on the semantic web. In *Proceedings of the 3th International Semantic Web Conference (ISWC)*, volume 3298 of *Lecture Notes in Computer Science*. Springer, 2004.
- [37] N. Guarino and C. A. Welty. A formal ontology of properties. In R. Dieng and O. Corby, editors, *Knowledge Acquisition, Modeling and Management, 12th International Conference, EKAW 2000, Juan-les-Pins, France, October 2-6, 2000*, pages 97–112, Berlin, 2000. Springer.
- [38] N. Guarino and C. A. Welty. An overview of ontoclean. In S. Staab and R. Studer, editors, *Handbook on Ontologies*, International Handbooks on Information Systems, chapter 8, pages 151–172. Springer, 2004.
- [39] I. Horrocks, P. F. Patel-Schneider, and F. van Harmelen. From shiq and rdf to owl: the making of a web ontology language. *J. Web Sem.*, 1(1):7–26, 2003.
- [40] H.R.Turtle. *Inference Networks for Document Retrieval*. PhD thesis, 1991.
- [41] D. Huynh, S. Mazzocchi, and D. Karger. Piggy bank: Experience the semantic web inside your web browser. In *International Semantic Web Conference (ISWC)*, 2005.
- [42] N. Ide and J. Veronis. Word sense disambiguation. *Special Issue on Computational Linguistics*, 24(1), 1998.

- [43] H. Knublauch, O. Dameron, and M. A. Musen. Weaving the biomedical semantic web with the prote'geowl plugin. In *KR-MED 2004, First International Workshop on Formal Biomedical Knowledge Representation, Proceedings of the KR 2004 Workshop on Formal Biomedical Knowledge Representation, Whistler, BC, Canada, 1 June 2004*, volume 102 of *CEUR Workshop Proceedings*, pages 39–47. CEUR-WS.org, 2004.
- [44] H. Knublauch, R. W. Fergerson, N. F. Noy, and M. A. Musen. The Protégé OWL plugin: An open development environment for semantic web applications. In *Lecture Notes in Computer Science*, volume 3298, pages 229–243, January 2004.
- [45] M. Kobayashi and K. Takeda. Information retrieval on the web. *ACM Computing Surveys*, 32(2):144–173, 2000.
- [46] J. Leukel, V. Schmitz, and F.-D. Dorloff. A modeling approach for product classification systems.
- [47] C. Masolo and S. Borgo. Qualities in formal ontology. In *Proceedings of the Ws Foundational Aspects of Ontologies (FOnt 2005)*, Koblenz, Germany, 2005.
- [48] C. Masolo, S. Borgo, A. Gangemi, N. Guarino, and A. Oltramari. Wonderweb deliverable d18. ontology library (final). Deliverable D18, WonderWeb. Ontology Infrastructure for the Semantic Web, 2006.12.31 2003.
- [49] C. Masolo, S. Borgo, A. Gangemi, N. Guarino, and A. Oltramari. *The WonderWeb Library of Foundational Ontologies (D18)*. Laboratory for Applied Ontology - ISTC-CNR, 2003.
- [50] C. Masolo, S. Borgo, A. Gangemi, N. Guarino, A. Oltramari, and L. Schneider. Wonderweb deliverable d17. the wonderweb library of foundational ontologies. preliminary report. Deliverable D17, WonderWeb. Ontology Infrastructure for the Semantic Web, 2006.05.29 2003.
- [51] C. Masolo, G. Guizzardi, L. Vieu, E. Bottazzi, and R. Ferrario. Relational roles and qua-individuals. In *Proceedings of the AAAI Fall Symposium on Roles, an interdisciplinary perspective*, 2005.
- [52] C. Masolo, L. Vieu, E. Bottazzi, C. Catenacci, R. Ferrario, A. Gangemi, and N. Guarino. Social roles and their descriptions. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Ninth International Conference (KR2004)*, Whistler, Canada, June 2-5, 2004, pages 267–277, 2004.

- [53] D. L. McGuinness and F. van Harmelen. OWL Web Ontology Language overview. Technical report, W3C, 2004.
- [54] M. Missikoff, R. Navigli, and P. Velardi. Integrated approach for web ontology learning and engineering. *IEEE Computer*, -:54–57, November 2002.
- [55] P. Mitra, G. Wiederhold, and M. L. Kersten. A graph-oriented model for articulation of ontology interdependencies. In C. Zaniolo, P. C. Lockemann, M. H. Scholl, and T. Grust, editors, *Advances in Database Technology - EDBT 2000, 7th International Conference on Extending Database Technology, Konstanz, Germany, March 27-31, 2000*, pages 86–100, 2000.
- [56] D. Nardi and R. J. Brachman. *An Introduction to Description Logics.*, pages 1–40. Cambridge University Press, Cambridge, 2003.
- [57] H. Papageorgiou, M. Vardaki, M. Petrakos, E. Theodorou, and F. Pentaris. Harmonisation of economic statistical classifications and related transformations. In *New Techniques and Technologies for Statistics (NTTS). Pre-Proceedings*, volume 1, pages 345–354, 2001.
- [58] H. Papageorugiou, M. Vardaki, M. Petrakos, E. Theodorou, and F. Pentaris. Harmonisation of economic statistical classifications and related transformations. Web.
- [59] S. E. Preece. *A spreading activation network model for information retrieval*. PhD thesis, 1981.
- [60] Y. Qiu and H.-P. Frei. Concept-based query expansion. In *Proceedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval*, pages 160–169, Pittsburgh, US, 1993.
- [61] D. Ravishankar, K. Thirunarayan, and T. Immaneni. A modular approach to document indexing and semantic search. In A. Press, editor, *Web Technologies, Applications, and Services*, pages 494–500, 2005.
- [62] A. L. Rector. Modularisation of domain ontologies implemented in description logics and related formalisms including owl. In *Proceedings of the 2nd International Conference on Knowledge Capture (K-CAP 2003), October 23-25, 2003, Sanibel Island, FL, USA*, pages 121–128, 2003.
- [63] C. Rocha, D. Schwabe, and M. P. de Aragão. A hybrid approach for searching in the semantic web. In *WWW*, pages 374–383, 2004.

- [64] C. Rocha, D. Schwabe, and M. P. de Aragão. Integrating semantic concept similarity in model-based web applications. In *WebMedia/LA-WEB*, pages 78–85, 2004.
- [65] U. Sattler. Description logics for the representation of aggregated objects. In W.Horn, editor, *Proceedings of the 14th European Conference on Artificial Intelligence*, Amsterdam, 2000. IOS Press.
- [66] D. Sperber and D. Wilson. *Relevance: Communication and cognition*. Harvard University Press and Oxford: Blackwell, 1986.
- [67] P. Vossen. Eurowordnet, general document. Technical report, University of Amsterdam, 1999.
- [68] C. A. Welty and N. Guarino. Supporting ontological analysis of taxonomic relationships. *Data Knowledge Engineering*, 39(1):51–74, 2001.