



UNIVERSIDADE DA CORUÑA

FACULTAD DE INFORMÁTICA

DEPARTAMENTO DE
TECNOLOGÍAS DE LA INFORMACIÓN Y LAS COMUNICACIONES

TESIS DOCTORAL

**UNA APROXIMACIÓN MULTI-CRITERIO PARA LA
SELECCIÓN AUTOMÁTICA DE ONTOLOGÍAS BIOMÉDICAS
UTILIZANDO CONOCIMIENTO COLECTIVO**

Marcos Martínez Romero

Directores:

Dr. Alejandro Pazos Sierra
Dr. José Manuel Vázquez Naya

A Coruña, noviembre 2010

The Web as I envisaged it, we have not seen it yet.

The future is still so much bigger than the past.

Sir Tim Berners-Lee

Agradecimientos

Varias son las personas a las que quiero agradecer el gran apoyo y ayuda prestados durante el transcurso de esta tesis.

En primer lugar, deseo expresar mi más sincero agradecimiento a dos personas que, además de los directores de este trabajo, han sido mis maestros y guías durante este camino. A Alejandro, por darme la oportunidad, ya cinco años atrás, de comenzar un proyecto fin de carrera que terminó convirtiéndose en la semilla de esta tesis. Por facilitarme todos los medios necesarios para llevarla a cabo. Por su apoyo y confianza. A José, por guiarme durante el camino con infinidad de sabios consejos. Por constituir una referencia para mí, en lo profesional y en lo personal. Por su valioso criterio crítico. Por animarme y depositar su confianza en mí. Por su amistad.

A Javi, por su continuo apoyo y confianza, por las oportunidades que me ha dado.

A Norberto, por sus sabios consejos científicos. Por brindarme su ayuda y amistad.

A todos los componentes del grupo RNASA-IMEDIR, por ser como una gran familia para mí. A Fran, por su amistad, compañerismo, apoyo y ayuda. A Marcos Gestal, por la ayuda prestada durante el tramo final del trabajo. A José Antonio, por su compañía durante la estancia en Aveiro.

También me gustaría mostrar mi agradecimiento a las personas que se han preocupado por mí durante mis estancias en el extranjero. A Paul Chapman, por acogerme durante mi estancia en la Universidad de Hull. Gracias también a José Luis Oliveira y a los miembros del Bioinformatics Group, por su hospitalidad durante mi paso por la Universidad de Aveiro.

Mi agradecimiento también está dirigido a todas aquellas personas que, con sus sugerencias y comentarios, me han ayudado a iniciar, avanzar y mejorar el trabajo. A Mark Musen, por compartir su conocimiento y experiencia conmigo. A Miriam Fernández, por las ideas para mejorar la aproximación. A los expertos que han cubierto el cuestionario de evaluación: Victoria Petri, Martin Krallinger, Yu Lin y Stefan Schulz. A Darren Natale, Barry Smith y Werner Ceusters, por haber revisado el prototipo y haber proporcionado sugerencias para mejorarlo. Gracias también a Juan Pazos, por sus consejos y correcciones.

Quiero también dar las gracias a todos los que, con su amistad y compañía en momentos de ocio, me han ayudado hacer más llevadero el camino. A David y a Yago, por estar ahí desde que tengo uso de razón y por continuar ocupando ese lugar, a pesar de la distancia. A Jacobo, por acompañarme a Cardiff a presentar una versión inicial de este trabajo. A Uiara y Denise, por su compañía durante mi estancia en Aveiro.

En cuanto a mi familia, no existen palabras para expresar lo agradecido que me siento por todo el cariño y apoyo incondicional que me ha dado. Quiero dar las gracias a mis padres, por ser la base y el impulso. Por haber sabido exigirme y premiarlo, enseñándome la importancia del esfuerzo y la constancia. A mi hermana Alba, por todos los momentos felices que hemos pasado juntos desde que éramos niños. Por preocuparse por mí en todo momento. A mi abuela Carmen y a mi tía Mari, por proporcionarme un segundo hogar y ayudarme en todo lo que he necesitado. También a mi abuelo Celso, por facilitarme los medios para introducirme en el mundo de la Informática. Me lo habéis dado todo, y siempre estaré en deuda con vosotros.

Finalmente, quiero expresar un especial agradecimiento a Sofía, por acompañarme y apoyarme desde el principio. Por ser la persona que más de cerca ha vivido la otra cara del esfuerzo realizado, y haberme dado el ánimo y el cariño necesarios para continuar siempre adelante. Por comprenderme. Por enseñarme que lo más importante no es llegar, sino el camino en sí mismo.

Muchas gracias a todos, de corazón.

Marcos.

a Sofía

Resumo

A día de hoxe, as ontoloxías considéranse unha ferramenta importante para estruturar e reutilizar a gran cantidade de información existente, especialmente en dominios como a biomedicina, nos que a axeitada organización e tratamento da información son aspectos críticos. Nestes dominios, o número de ontoloxías dispoñibles creceu moi rapidamente durante os últimos anos. Este feito, que é moi positivo, pois permite un manexo máis eficiente (ou intelixente) da información, formula un novo problema: ¿que ontoloxía utilizar para unha tarefa determinada?

Reutilizar as ontoloxías existentes en lugar de crear ontoloxías novas é unha práctica deseable. Construír unha ontoloxía dende cero é unha tarefa moi complexa e que require moito tempo e recursos humanos especializados. Ademais, para asegurar unha axeitada interoperabilidade, é necesario evitar a existencia de múltiples ontoloxías que representan o mesmo coñecemento. Non obstante, a causa do crecente número, complexidade e variedade de ontoloxías existentes, formadas en moitos casos por miles de conceptos e relacóns entre eles, elixir a ontoloxía ou ontoloxías para reutilizar nun problema de anotación semántica ou para deseñar unha aplicación específica é unha tarefa difícil. Debido a isto, o desenvolvemento de aproximacións e ferramentas que faciliten a selección da mellor ontoloxía ou ontoloxías a utilizar nun contexto determinado, estase a converter nunha prioridade para os investigadores.

Os criterios a ter en conta para seleccionar unha ontoloxía para unha tarefa son moitos: o número de conceptos e relacóns que contén, a situación dos conceptos na ontoloxía, a linguaxe na que se encontra representada, etc. Non obstante, é necesario elixir os criterios más relevantes e combinalos de forma precisa, para facer posible a obtención de bos resultados nun período de tempo razonable e sen necesidade de intervención dun experto.

Neste traballo, proponse unha aproximación para a selección automática de ontoloxías biomédicas, que se basea en medir a adecuación dunha ontoloxía a un contexto determinado de acordo a tres criterios: (1) O grao en que a ontoloxía cubre o contexto. (2) A riqueza semántica da ontoloxía no contexto. (3) A popularidade da ontoloxía na comunidade biomédica.

Para validar a aproximación, implementouse un prototipo de sistema de selección de ontoloxías no dominio biomédico, que se avaliou en varios escenarios habituais de reutilización de ontoloxías en biomedicina.

Resumen

A día de hoy, las ontologías se consideran una herramienta importante para estructurar y reutilizar la gran cantidad de información existente, especialmente en dominios como la biomedicina, en los que la adecuada organización y tratamiento de la información son aspectos críticos. En estos dominios, el número de ontologías disponibles ha crecido muy rápidamente durante los últimos años. Este hecho, que es muy positivo, pues permite un manejo más eficiente (o inteligente) de la información, plantea un nuevo problema: ¿qué ontología utilizar para una tarea determinada?

Reutilizar las ontologías existentes en lugar de crear ontologías nuevas es una práctica deseable. Construir una ontología desde cero es una tarea muy compleja y que requiere mucho tiempo y recursos humanos especializados. Además, para asegurar una adecuada interoperabilidad, es necesario evitar la existencia de múltiples ontologías que representan el mismo conocimiento. Sin embargo, a causa del creciente número, complejidad y variedad de ontologías existentes, formadas en muchos casos por miles de conceptos y relaciones entre ellos, elegir la ontología u ontologías para reutilizar en un problema de anotación semántica o para diseñar una aplicación específica es una tarea difícil. Debido a esto, el desarrollo de aproximaciones y herramientas que faciliten la selección de la mejor ontología u ontologías a utilizar en un contexto determinado, se está convirtiendo en una prioridad para los investigadores.

Los criterios a tener en cuenta para seleccionar una ontología para una tarea son muchos: el número de conceptos y relaciones que contiene, la ubicación de los conceptos en la ontología, el lenguaje en el que se encuentra representada, etc. Sin embargo, es necesario elegir los criterios más relevantes y combinarlos de forma precisa, para hacer posible la obtención de buenos resultados en un período de tiempo razonable y sin necesidad de intervención de un experto.

En este trabajo, se propone una aproximación para la selección automática de ontologías biomédicas, que se basa en medir la adecuación de una ontología a un contexto determinado de acuerdo a tres criterios: (1) El grado en que la ontología cubre el contexto. (2) La riqueza semántica de la ontología en el contexto. (3) La popularidad de la ontología en la comunidad biomédica.

Para validar la aproximación, se ha implementado un prototipo de sistema de selección de ontologías en el dominio biomédico, que se ha evaluado en varios escenarios habituales de reutilización de ontologías en biomedicina.

Abstract

Nowadays, ontologies are considered an important tool for structuring and reusing the vast amount of information, especially in domains such as biomedicine, in which the proper organization and processing of information are critical issues. In these domains, the number of available ontologies has grown rapidly during the last years. This is very positive, because it enables a more efficient (or more intelligent) knowledge management. However, it raises a new problem: what ontology should be used for a given task?

Reusing existing ontologies rather than creating new ones is a desirable practice. Building an ontology from scratch is a very complex and time-consuming task, which requires specialized human resources. In addition, avoiding the existence of multiple ontologies that represent the same knowledge is necessary to ensure proper interoperability. However, due to the increasing number, complexity and variety of existing ontologies, frequently containing thousands of concepts and relations between them, choosing the ontology or ontologies to be reused in a semantic annotation problem or to design a specific application is a difficult task. Due to this, the development of approaches and tools that facilitate the task of selecting the best ontology or ontologies for a given context is becoming a priority for researchers.

Many criteria may be taken into account when selecting an ontology for a given task: the number of concepts and relationships that it contains, the location of concepts in the ontology, the language in which it is represented, etc. However, it is necessary to select the most relevant criteria and combine them accurately, in order to make it possible to obtain good results in a reasonable time and without intervention of an expert.

In this work, an approach for the automatic selection of biomedical ontologies is proposed. This approach is based on measuring the adequacy of an ontology to a given context according to three criteria: (1) The extent to which the ontology covers the context. (2) The semantic richness of the ontology in the context. (3) The popularity of the ontology in the biomedical community.

In order to validate the approach, a system for the selection of ontologies in the biomedical domain has been prototyped, which has been evaluated in several typical knowledge reusing scenarios in biomedicine.

Índice de Contenido

1	INTRODUCCIÓN.....	1
1.1	CONTEXTUALIZACIÓN.....	1
1.2	PLANTEAMIENTO DEL PROBLEMA	4
1.3	OBJETIVOS.....	7
1.4	HIPÓTESIS.....	8
1.5	ORGANIZACIÓN DE LA MEMORIA	8
2	FUNDAMENTOS.....	11
2.1	ONTOLOGÍAS	11
2.1.1	<i>Definición</i>	14
2.1.2	<i>Típos de ontologías.....</i>	17
2.1.3	<i>Lenguajes de representación de ontologías</i>	20
2.1.4	<i>Aplicaciones de las ontologías.....</i>	24
2.1.5	<i>Bio-ontologías</i>	25
2.1.6	<i>Repositorios de ontologías.....</i>	38
2.1.7	<i>Evaluación y selección de ontologías</i>	42
2.2	WEB 2.0 Y CONOCIMIENTO COLECTIVO.....	43
2.2.1	<i>De la Web 1.0 a la Web 2.0</i>	43
2.2.2	<i>Conocimiento colectivo</i>	47
2.3	LA WEB SEMÁNTICA	49
2.3.1	<i>¿Por qué es necesaria una Web Semántica?</i>	52
2.3.2	<i>Arquitectura de la Web Semántica</i>	55
2.3.3	<i>El papel de las ontologías en la Web Semántica.....</i>	58
2.3.4	<i>Web Semántica vs. Web 2.0.....</i>	59
2.4	METADATOS Y ANOTACIÓN SEMÁNTICA	61
2.4.1	<i>Metadatos y metadatos semánticos</i>	61
2.4.2	<i>Calidad de los datos y metadatos semánticos</i>	63
2.4.3	<i>¿Qué es la anotación?</i>	64
2.4.4	<i>Anotación semántica.....</i>	66
2.4.5	<i>Beneficios de la anotación semántica.....</i>	70
3	ESTADO DE LA CUESTIÓN.....	73
3.1	TERMINOLOGÍA Y DEFINICIONES	74
3.1.1	<i>Evaluación técnica y evaluación de usuario.....</i>	75
3.1.2	<i>Ordenación, recomendación y selección</i>	78
3.2	CARACTERIZACIÓN DE LAS ONTOLOGÍAS Y CRITERIOS DE EVALUACIÓN	79
3.2.1	<i>Criterios de evaluación de ontologías biomédicas.....</i>	83
3.3	REQUERIMIENTOS DE LA SELECCIÓN DE ONTOLOGÍAS	85

3.3.1	<i>Selección manual vs. automática</i>	85
3.4	APROXIMACIONES EXISTENTES	86
3.4.1	<i>Evaluación para el desarrollo de ontologías</i>	89
3.4.2	<i>Evaluación para la selección de ontologías</i>	94
3.4.3	<i>Comparativa de aproximaciones</i>	109
3.4.4	<i>Limitaciones de las actuales aproximaciones</i>	111
4	MÉTODOS	115
5	APROXIMACIÓN PROPUESTA	117
5.1	DESCRIPCIÓN GENERAL.....	117
5.2	EXPANSIÓN SEMÁNTICA	119
5.2.1	<i>Normalización y corrección</i>	120
5.2.2	<i>Identificación de conceptos</i>	121
5.2.3	<i>Desambiguación</i>	122
5.2.4	<i>Expansión semántica</i>	133
5.3	RECUPERACIÓN DE ONTOLOGÍAS.....	134
5.4	EVALUACIÓN DE ONTOLOGÍAS.....	137
5.4.1	<i>Evaluación de la cobertura del contexto</i>	137
5.4.2	<i>Evaluación de la riqueza semántica</i>	141
5.4.3	<i>Evaluación de la popularidad</i>	148
5.4.4	<i>Agregación de puntuaciones</i>	150
5.4.5	<i>Normalización de valores</i>	151
5.4.6	<i>Ajuste de pesos</i>	153
5.5	COMBINACIÓN Y ORDENACIÓN DE ONTOLOGÍAS.....	155
5.6	DIFERENCIAS CON LAS APROXIMACIONES EXISTENTES.....	157
6	IMPLEMENTACIÓN DE LA APROXIMACIÓN PROPUESTA	161
6.1	DESCRIPCIÓN GENERAL	161
6.1.1	<i>Estructura de paquetes</i>	162
6.2	EXPANSIÓN SEMÁNTICA	165
6.2.1	<i>Normalización y corrección</i>	165
6.2.2	<i>Identificación de conceptos</i>	167
6.2.3	<i>Desambiguación</i>	169
6.2.4	<i>Expansión semántica</i>	172
6.3	RECUPERACIÓN DE ONTOLOGÍAS.....	173
6.3.1	<i>Construcción de un repositorio de ontologías biomédicas</i>	174
6.3.2	<i>Almacenamiento y acceso a las ontologías</i>	179
6.4	EVALUACIÓN DE ONTOLOGÍAS.....	182
6.4.1	<i>Evaluación de la cobertura del contexto</i>	183

6.4.2	<i>Evaluación de la riqueza semántica</i>	186
6.4.3	<i>Evaluación de la popularidad</i>	191
6.4.4	<i>Agregación de puntuaciones</i>	194
6.4.5	<i>Normalización de valores</i>	194
6.4.6	<i>Ajuste de pesos</i>	197
6.5	COMBINACIÓN Y ORDENACIÓN DE ONTOLOGÍAS.....	198
6.6	INTERFACES PROPORCIONADAS.....	201
6.6.1	<i>Servicio Web</i>	201
6.6.2	<i>Interfaz gráfica</i>	202
7	EVALUACIÓN	205
7.1	DESCRIPCIÓN DEL EXPERIMENTO	205
7.2	RESULTADOS PROPORCIONADOS POR EL PROTOTIPO	208
7.2.1	<i>Caso de estudio 1 (cáncer de mama)</i>	209
7.2.2	<i>Caso de estudio 2 (cuidados críticos)</i>	210
7.2.3	<i>Caso de estudio 3 (registro de cáncer)</i>	211
7.2.4	<i>Caso de estudio 4 (anatomía)</i>	214
7.2.5	<i>Caso de estudio 5 (ontología de ejemplo)</i>	216
7.2.6	<i>Tiempos de ejecución</i>	218
7.3	RESULTADOS DE LA EVALUACIÓN	220
8	CONCLUSIONES Y FUTURAS LÍNEAS DE INVESTIGACIÓN.....	227
8.1	CONCLUSIONES.....	227
8.2	FUTURAS LÍNEAS DE INVESTIGACIÓN	229
9	CONCLUSIONS AND FUTURE RESEARCH LINES	231
9.1	CONCLUSIONS.....	231
9.2	FUTURE RESEARCH LINES.....	233
	ACRÓNIMOS.....	235
	ANEXO I. ONTOLOGÍAS UTILIZADAS.....	239
	ANEXO II. FORMULARIOS.....	243
	ANEXO III. INTERFAZ DEL SERVICIO WEB	257
	REFERENCIAS	263

1 Introducción

En este primer capítulo, se introduce brevemente el contexto en el que se enmarca la tesis, para describir a continuación el problema que se pretende afrontar y poder plantear los objetivos del trabajo y la hipótesis en la que se centra la investigación realizada. El último apartado de este capítulo se dedica a explicar la organización del resto del documento y el propósito de cada uno de los capítulos restantes.

1.1 Contextualización

Durante las últimas décadas, e impulsados por iniciativas como el ARPA Knowledge Sharing Effort (Neches et al., 1991), se han realizado importantes progresos en la concepción e implantación de una infraestructura tecnológica orientada a facilitar la compartición y reutilización del conocimiento entre diferentes sistemas inteligentes. La finalidad de este esfuerzo es facilitar la construcción de sistemas mayores y más potentes, constituidos por componentes reutilizables, capaces de proporcionar nuevos avances en diversos dominios a partir de una gestión adecuada de la inmensa cantidad de información disponible actualmente en formato digital.

Para esto, los enfoques de tratamiento de información tradicionales, basados en el “procesamiento de datos”, se están trasladando hacia una perspectiva de “procesamiento de conceptos”, que permita a los computadores tratar grandes cantidades de información en base a su semántica, de forma similar a como lo harían los seres humanos pero de forma más rápida y a una escala mayor. Este nuevo enfoque requiere disponer de estrategias formales para describir los datos semánticamente, de

tal manera que éstos puedan ser “entendidos” por las máquinas. Para afrontar este problema, surgen las ontologías.

Desde principios de los 90 y, en especial, desde la concepción de la Web Semántica (Berners-Lee et al., 2001), las ontologías han ido despertando cada vez más interés en el ámbito de las Ciencias de la Computación. En la actualidad, las ontologías son la estructura más frecuentemente utilizada para representar conceptos y relaciones entre conceptos, en un dominio particular (Brank et al., 2005). Aunque existen múltiples definiciones de ontología, ligeramente diferentes, una ontología se suele definir como una “*especificación formal y explícita de una conceptualización compartida*” (Studer et al., 1998). Las ontologías constituyen la base para permitir la compartición y reutilización de conocimiento, pues establecen vocabularios e interpretaciones semánticas comunes (Gómez-Pérez, 1995). Además, es importante resaltar que el conocimiento representado utilizando ontologías puede ser procesado por los computadores, haciendo posible automatizar tareas de razonamiento, búsqueda o análisis de datos a gran escala. Actualmente, las ontologías se están utilizando en aplicaciones relacionadas con la Gestión del Conocimiento, Bioinformática y Biomedicina, Procesamiento del Lenguaje Natural, Integración Inteligente de Información o Comercio Electrónico, etc., principalmente en tareas de anotación e integración de datos y construcción de aplicaciones inteligentes.

Debido a los importantes beneficios potenciales derivados de disponer de información descrita de forma semántica mediante ontologías, durante la última década se han desarrollado miles¹ de ontologías de múltiples dominios. La gran mayoría de estas ontologías se encuentran accesibles libremente a través de Internet, bien publicadas de forma aislada, bien almacenadas en repositorios de ontologías, y este número continúa creciendo día a día.

El desarrollo de ontologías es una tarea compleja (Alani et al., 2007), que requiere considerables inversiones de tiempo y recursos, además de expertos con un gran conocimiento del dominio en cuestión. Debido a esto y para conseguir una adecuada interoperabilidad, la reutilización de ontologías es clave. Supóngase, por ejemplo, que se desea desarrollar una aplicación Web que se basa en una ontología. Reutilizar una

¹ Actualmente, el buscador semántico Swoogle indexa más de 10.000 ontologías.

ontología ya existente permite una interoperabilidad directa con otros sistemas que ya se basan en esa ontología, además de ahorrar la gran cantidad de tiempo y dinero necesarios para desarrollar una ontología adecuadamente consensuada y testeada.

Sin embargo, el gran número de ontologías existentes actualmente dificulta de forma considerable su reutilización. Encontrar la ontología u ontologías más adecuadas para describir un contexto particular es una tarea muy compleja, que consume una gran cantidad de tiempo (cada vez mayor, a medida que el número de ontologías aumenta) y requiere disponer de un elevado conocimiento del dominio en cuestión.

Debido a esto, a medida que el número y diversidad de ontologías continúa creciendo, también aumenta la necesidad de disponer de estrategias y herramientas que simplifiquen el proceso de identificar cuál es la ontología u ontologías más apropiadas para describir un dominio, contexto o tarea particular. Es sabido que una buena ontología es aquélla que satisface el propósito para el que fue construida (Brewster et al., 2004), pero no resulta factible asumir que los usuarios serán capaces y dispondrán del tiempo necesario para seleccionar la ontología u ontologías más adecuadas para un determinado propósito.

También hay que tener en cuenta que elegir una ontología que no satisface adecuadamente las necesidades del usuario o del sistema que la va a utilizar, puede obligar a usuarios futuros a dejar de utilizar dicha ontología y forzarlos a formalizar el mismo conocimiento de nuevo (Arpírez et al., 2000). Para soportar la reutilización de conocimiento a gran escala, en entornos abiertos como la actual Web Semántica, es crucial disponer de mecanismos robustos para la selección de ontologías (Sabou et al., 2006a). La reutilización de ontologías no sólo permitirá alcanzar una mejor interoperabilidad entre aplicaciones y sistemas, sino que también liberará a los ingenieros de conocimiento de la tarea de construir ontologías desde cero (Lewen et al., 2006). Como indicaba Asunción Gómez-Pérez en 1996, “*juzgar una ontología antes de su uso para representar conocimiento es simplemente tan necesario como realizar una revisión a fondo de un coche de segunda mano antes de comprarlo*” (Gómez-Pérez, 1996). Estos mecanismos deben realizar una evaluación exhaustiva de cada ontología, teniendo en cuenta diversos aspectos, como por ejemplo el propósito de la ontología, la madurez de su contenido, el nivel de actualizaciones, su relevancia en la comunidad, etc. Toda esta información subjetiva es crítica para determinar si una ontología es adecuada o no para

un propósito concreto (Lewen, et al., 2006). Sin embargo, obtenerla y evaluarla de forma objetiva es una tarea muy compleja.

Uno de los campos en el que las ontologías han evolucionado más rápidamente y en el que se pueden encontrar más ejemplos de ontologías es la biomedicina. Las ontologías biomédicas son ampliamente utilizadas para diseñar sistemas de recuperación de información, para facilitar la interoperabilidad entre repositorios de datos, y para desarrollar sistemas que parsean, anotan o indexan recursos de datos biomédicos. En biomedicina, los investigadores utilizan ontologías para anotar (o etiquetar) semánticamente sus datos (Bodenreider & Stevens, 2006), facilitando así su integración con otros datos y el descubrimiento de conocimiento nuevo.

1.2 Planteamiento del problema

Durante los últimos años, han surgido varias aproximaciones orientadas a facilitar la selección de la ontología u ontologías más adecuadas para describir un contexto determinado. Sin embargo, estas aproximaciones presentan algunos inconvenientes que se pueden considerar un impedimento para la adecuada reutilización de ontologías. Entre estos problemas, se pueden destacar los siguientes:

- Muchas de las aproximaciones propuestas hasta el momento **requieren de intervención por parte del usuario**. No son completamente automáticas, lo cual impide su uso en entornos que requieren de reutilización automática de conocimiento (e.g. Web Semántica).
- La información de entrada proporcionada para llevar a cabo el proceso de selección de ontologías se limita a **una palabra clave**. Esto supone una restricción importante si se desea, por ejemplo, seleccionar la mejor ontología para describir semánticamente un conjunto de términos que representan un contexto.
- **Evaluación inadecuada de la popularidad.** Por definición, una ontología contiene conocimiento compartido, o consensuado. Teniendo esto en cuenta, cualquier método diseñado para evaluar una ontología debería tener en cuenta

el nivel de aceptación colectiva, o “popularidad”², de la ontología. Algunas de las aproximaciones existentes tienen en cuenta este criterio. Sin embargo, se limitan a medir dicha popularidad según los enlaces inter-ontología, asumiendo que *“la relevancia de una ontología es proporcional al número de ontologías que la referencian”*. Varios autores han explicado que esta forma de medir la popularidad de ontologías, ideada a partir del conocido PageRank de Google (Page et al., 1998), es problemática y no siempre proporciona buenos resultados (Sabou et al., 2006b). Otras aproximaciones calculan la popularidad en base a información proporcionada de forma directa por los usuarios. Esta idea tampoco es adecuada en un entorno real. No se puede pensar que los usuarios estarán dispuestos a dedicar su tiempo a valorar las ontologías.

- **No disponen de salida combinada.** En general, cuando se desea describir semánticamente grandes conjuntos de términos, es poco frecuente encontrar una única ontología que cubra todos ellos. En estos casos, resultaría útil saber cómo se comportarían diferentes combinaciones de ontologías. Sin embargo, las actuales aproximaciones se limitan a seleccionar una única ontología.
- Otro problema de las aproximaciones de selección de ontologías existentes es que **no consideran a las ontologías como artefactos de conocimiento**. Son tratadas como un conjunto de objetos interconectados, grafos o colecciones de cadenas de texto, pero no se explota su potencial semántico, ni tampoco el pragmático. El significado de los términos de búsqueda y los conceptos identificados en las ontologías es ignorado por completo. Las estrategias de selección deberían ser mejoradas hacia una evaluación más semántica, que tenga en cuenta el sentido y contexto de los términos de búsqueda (Sabou, et al., 2006b).
- Otro inconveniente es el **bajo rendimiento** de algunas aproximaciones, debido a que se basan en la aplicación de métricas de gran complejidad. Un ejemplo es AKTiveRank. Como explican Sabou y colegas (Sabou, et al., 2006b), este sistema requiere de aproximadamente dos minutos para evaluar cada ontología. Esto es un gran impedimento a la hora de trabajar con repositorios que

² La 22^a edición del diccionario de la RAE define popularidad como “Aceptación y aplauso que alguien tiene en el pueblo”.

contienen cientos de ontologías, en entornos que requieren de una respuesta rápida. Mantener un equilibrio entre la complejidad de los métodos de selección utilizados y el rendimiento, es un aspecto crucial.

Debido a estos inconvenientes, se puede afirmar que a día de hoy no existe ninguna aproximación para la selección de ontologías que proporcione una recomendación clara acerca de qué ontología u ontologías son más adecuadas para describir un contexto determinado, y que sea válida para entornos en los que se requiera reutilización automática de conocimiento.

En esta tesis, se propone una aproximación para la selección automática de ontologías, dirigida a facilitar la tarea de seleccionar la mejor ontología u ontologías para representar semánticamente un contexto particular. Para esto, en esta aproximación se combinan tres aspectos ontológicos distintos (ver figura 1.1):

1. **El grado de cobertura del contexto** proporcionado por cada ontología.
2. **La riqueza del conocimiento** expresado por cada ontología en el contexto.
3. **La popularidad** de cada ontología en la Web 2.0.

La combinación de estos tres métodos de evaluación, está orientada a proporcionar una selección automática y de calidad en un tiempo razonable, que pueda ser de utilidad tanto para usuarios como para agentes software, que necesiten afrontar tareas de reutilización automática de conocimiento en entornos como la Web Semántica.

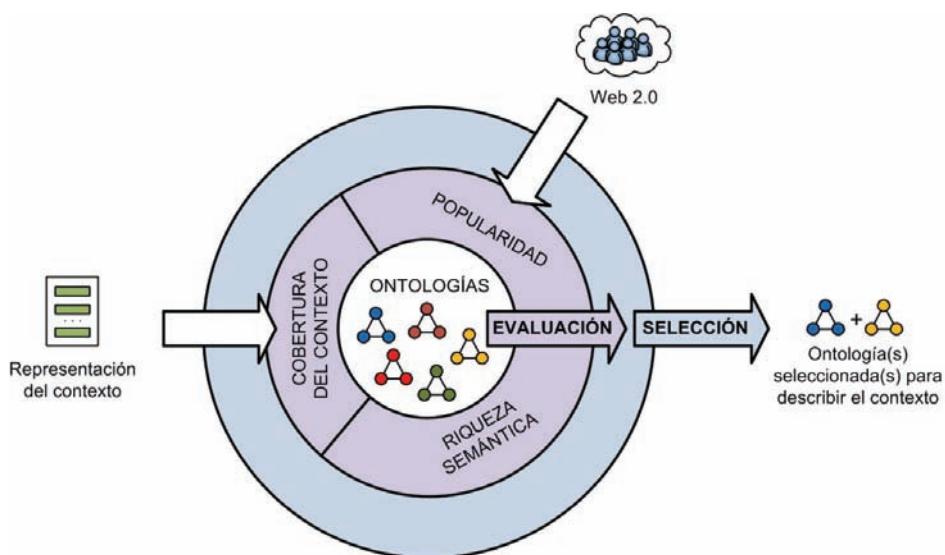


Figura 1.1. Esquema general de los criterios de evaluación en los que se basa la aproximación propuesta.

El problema de selección de ontologías que se afronta en esta tesis se puede dividir en tres subproblemas, diferentes pero complementarios: búsqueda de ontologías (*ontology search*), evaluación de ontologías (*ontology evaluation*) y selección de ontologías (*ontology selection*).

La aproximación se centrará en el dominio de la biomedicina. Como explican Alani y colegas (Alani, et al., 2007), centrarse en la recomendación de ontologías que cubren dominios similares o bien aspectos relacionados de un mismo dominio, supone un reto mucho mayor que trabajar en la recomendación de ontologías de dominios diferentes. A modo de ejemplo, supóngase que se desea seleccionar una ontología para describir un conjunto de términos del dominio del cáncer. En este caso, resultará más sencillo decidirse entre una ontología de medios de transporte y una ontología biomédica, que elegir entre dos ontologías biomédicas que pueden solaparse (e.g. una ontología clínica y otra sobre anatomía).

1.3 Objetivos

El principal objetivo de esta tesis consiste en idear una aproximación para la selección de ontologías que permita proporcionar una recomendación acerca de cuál o cuáles son las ontologías más adecuadas para representar un contexto determinado. Esta aproximación:

1. Será automática. No requerirá intervención por parte del usuario durante el proceso de selección.
2. Para un conjunto de ontologías determinado, proporcionará la ontología u ontologías más adecuadas para describir el contexto representado por un conjunto de palabras clave.
3. Estará diseñada para poder ser integrada fácilmente con otros sistemas que requieran de funcionalidades de reutilización automática de conocimiento.
4. Se basará en la evaluación de la ontología de acuerdo a tres criterios diferentes, aunque complementarios: (1) El grado de cobertura del contexto. (2) La riqueza semántica de la ontología en el contexto. (3) La popularidad de la ontología.

5. Será capaz de medir la popularidad de cada ontología sin requerir la intervención del usuario.
6. Será capaz de proporcionar como salida combinaciones de ontologías.
7. Mantendrá un equilibrio adecuado entre la complejidad de los métodos de evaluación y el rendimiento de la aproximación.
8. Se centrará en el ámbito de las ontologías biomédicas.

1.4 Hipótesis

La hipótesis subyacente a este trabajo se puede enunciar de la forma siguiente:

“Combinando la evaluación del grado de cobertura que una ontología biomédica proporciona sobre un contexto determinado, la riqueza semántica de la ontología en dicho contexto y la popularidad de la ontología en la Web, es posible plantear una aproximación que permita seleccionar automáticamente la ontología u ontologías biomédicas más adecuadas para describir semánticamente dicho contexto”.

1.5 Organización de la memoria

Esta memoria se ha organizado en los siguientes capítulos:

- **Capítulo 1: Introducción.** En el presente capítulo, se realiza una introducción al problema que se pretende resolver: la selección automática de ontologías. Se lleva a cabo el planteamiento de problema y se identifican las principales carencias de las soluciones existentes. Se fijan los objetivos y se formula la hipótesis de trabajo.
- **Capítulo 2: Fundamentos.** Se explican la nociones fundamentales para comprender resto del contenido de la tesis. Se tratan los aspectos básicos de la representación de conocimiento usando ontologías, haciendo énfasis en su aplicación al dominio biomédico. También se exponen los aspectos más relevantes de la Web 2.0, la Web Semántica y del campo de la anotación semántica.

- **Capítulo 3: Estado de la cuestión.** En este capítulo se analizan detenidamente las soluciones existentes en los campos de la evaluación y selección de ontologías, resumiendo sus principales características y limitaciones.
- **Capítulo 4: Métodos.** Se describe la forma de proceder para resolver el problema planteado y poner a prueba la hipótesis de trabajo.
- **Capítulo 5: Aproximación propuesta.** Se expone la aproximación para la selección automática de ontologías biomédicas, que constituye el núcleo de esta tesis doctoral. En primer lugar, se describe de forma general el proceso de selección y los principales componentes que la conforman. A continuación, se explica cada uno de estos aspectos de forma más detallada, justificando las decisiones adoptadas y haciendo uso de diversos ejemplos para facilitar la comprensión. Al final del capítulo, se presentan las diferencias más relevantes de la aproximación con los trabajos existentes.
- **Capítulo 6: Implementación de la aproximación propuesta.** Se proporcionan los detalles de implementación de un prototipo de sistema de selección de ontologías biomédicas, que se ha desarrollado siguiendo la aproximación propuesta.
- **Capítulo 7: Evaluación.** Se realiza una evaluación de la aproximación propuesta. Se describen los casos de estudio utilizados y se analizan los resultados obtenidos.
- **Capítulo 8: Conclusiones y futuras líneas de investigación.** Se exponen las conclusiones derivadas del trabajo, y se plantean una serie de futuras líneas de investigación relacionadas con el mismo.
- **Capítulo 9: Conclusions and future research lines.** Constituye la traducción al inglés del capítulo 8, de acuerdo a lo exigido en la normativa vigente sobre la mención europea en el título de Doctor.

Se finaliza con la lista de acrónimos utilizados en el documento, los anexos y las referencias bibliográficas.

2 Fundamentos

En este capítulo, se explican en detalle las nociones necesarias para comprender el contenido del trabajo. Se comienza exponiendo los conceptos básicos de la representación del conocimiento utilizando ontologías, con especial atención al ámbito de las ontologías en biomedicina. A continuación, se explica en qué consisten la Web 2.0 y la Web Semántica, y cuáles son sus diferencias. Se finaliza el capítulo describiendo los aspectos fundamentales sobre metadatos y anotación semántica.

2.1 Ontologías

En Filosofía, Ontología es la ciencia de lo que es, de los tipos y estructuras de los objetos y sus propiedades, eventos, procesos y relaciones en todas las áreas de la realidad. Como explican Smith y Welty (Smith & Welty, 2001), esta Ontología adopta muchas formas, desde la Metafísica de Aristóteles (384-322 a. C.) (Aristóteles, 350 a. C.) hasta la Teoría de los Objetos de Alexius Meinong (1853-1920) (Meinong, 1904).

En la antigua Grecia, las taxonomías derivadas de la Ontología trataban de dar respuesta a preguntas como: ¿qué clases de entidades son necesarias para una descripción y explicación completa de todo lo que ocurre en el Universo?, o ¿qué clases de entidades se necesitan para dar cuenta de lo que hace verdad todas las verdades? Estas taxonomías se habían diseñado para ser exhaustivas en el sentido de que deberían incluir todos los tipos posibles de entidades, así como los tipos de relaciones entre ellas. Muchas de las ideas de la Ontología, como las nociones de *categoría* y *jerarquía*, se remontan a esta época (ver figura 2.1).

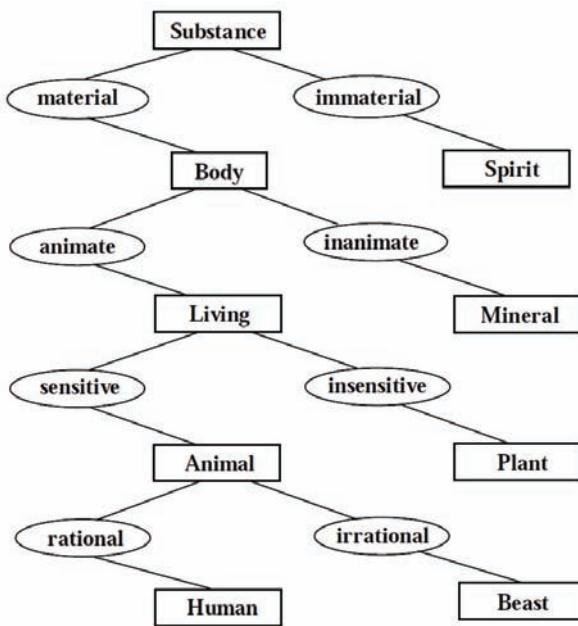


Figura 2.1. Árbol de Porfirio, con las categorías de Aristóteles (en rectángulos). Las líneas representan relaciones *is-a* (subsunción) entre categorías. Adaptado de (Sowa, 2000).

Durante los 2000 años posteriores a Aristóteles, el desarrollo la Ontología se estancó. Sin embargo, con la revolución científica (1500-1700 aprox.), las taxonomías filosóficas empezaron a sufrir avances radicales en cuanto a la comprensión del Universo, y las clasificaciones ya aceptadas de forma general comenzaron a cambiar. Fue en este momento cuando nace el término “ontología”, que fue acuñado en el año 1613 de forma independiente por dos filósofos, Rudolf Göckel (Goclenius), en su *Lexicon philosophicum* (Goclenius, 1613) y Jacob Lorhard (Lorhardus), en su *Theatrum philosophicum* (Lorhard, 1613).

De la misma forma que las raíces de la Ontología se entrelazaron con el desarrollo de la Filosofía y crecieron con ella, en los últimos años la Ontología se ha mezclado con el desarrollo de Inteligencia Artificial (IA) y de los Sistemas de Información. Desde un primer momento, la IA *logicista* centró su atención en sistemas que “conocen”, o son capaces de simular conocimiento, mediante el uso de mecanismos de razonamiento automático. A medida que estos mecanismos se estandarizaron con el paso del tiempo, se pasó a centrar la atención en sus teorías, dando lugar al nacimiento de la Gestión del Conocimiento (Del Moral et al., 2007). Paralelamente, expertos de otras áreas de la Informática comenzaron a detectar problemas a la hora de representar el conocimiento práctico o del mundo real. Así, en los Sistemas de Administración de

Bases de Datos se detectó el problema del *modelado conceptual*, y en la Ingeniería del Software se empezó a reconocer la importancia de lo que se entonces conoció como *modelado del dominio*.

En este contexto, tanto el ingeniero del conocimiento, como el modelador conceptual o el modelador del dominio, comenzaron a descubrir la necesidad de disponer de representaciones declarativas, que deberían ser lo suficientemente generales como para asegurar su reutilización, además de corresponderse con los objetos y procesos que representan. Así, surgieron preguntas como: ¿Qué es un objeto, un proceso, un atributo o una relación? ¿Qué es una transacción, una persona o una organización? ¿Cuáles son las dependencias entre ellos? ¿Cómo están relacionados?

En este punto, el paso hacia la Ontología en la IA se llevó a cabo por parte de individuos independientes de cada una de estas áreas. Fue John McCarthy el primero en reconocer el solapamiento entre el trabajo realizado en la Ontología filosófica y la actividad de construcción de teorías lógicas de los sistemas de IA. Ya en 1980, McCarthy afirmaba que los desarrolladores de sistemas inteligentes basados en lógica debían, en primer lugar, “*listar todo lo que existe, construir una ontología de nuestro mundo*” (McCarthy, 1980). Esta visión de McCarthy, inspirada en sus lecturas de Quine (Quine, 1968), fue asumida por Patrick Hayes en su trabajo sobre física *naïve* (Hayes, 1985). La mayoría de los esfuerzos de la Inteligencia Artificial en el campo logicista se centraron en capturar información sobre el mundo, compatible con la perspectiva del sentido común humano. Una perspectiva similar, pero con ambiciones mayores y con un reconocimiento más explícito del solapamiento con la filosofía, fue propuesta por John Sowa, quien se refiere a ontología como un mundo posible, un catálogo de todo lo que compone ese mundo, cómo se combina, y cómo funciona (Sowa, 1984).

A partir de los primeros usos del término “ontología” en el campo de la IA, como por ejemplo en Alexander y colegas (Alexander et al., 1986), el significado del término creció, y de la misma forma que los campos de Ingeniería de Conocimiento, Modelado Conceptual y Modelado del Dominio comenzaron a converger y a descubrirse mutuamente, también lo hicieron las diferentes variantes del significado de “ontología”, dando lugar a diversas definiciones, de diferentes autores. En el siguiente apartado, se repasan las principales definiciones del término “ontología” en el campo de la IA.

2.1.1 Definición

La palabra ontología (del griego *ontos=ser* y *logos=conocimiento*) tiene dos acepciones. Una de ellas, la más antigua, proviene de la Filosofía y es la que se refiere a la *explicación sistemática del ser*. Esta acepción es la que se recoge en el diccionario de la Real Academia Española (Real Academia Española, 2001): “*f. Parte de la metafísica que trata del ser en general y de sus propiedades trascendentales*”. La otra acepción surge a principios de los 90, dentro del campo de la Inteligencia Artificial. A continuación se recopilan las principales definiciones de la palabra ontología dentro del ámbito de la Inteligencia Artificial.

Una de las primeras definiciones fue la de Neches y colegas quienes, en 1991, definieron una ontología de la siguiente manera (Neches, et al., 1991):

“*Una ontología define los términos básicos y las relaciones que constituyen el vocabulario de un área específica, así como las reglas para combinar términos y relaciones para definir extensiones al vocabulario*”.

Esta definición descriptiva dice qué hacer para construir una ontología, y proporciona algunas guías a seguir: identifica términos básicos y relaciones entre términos, identifica reglas para combinar términos, y proporciona las definiciones de dichos términos y relaciones. Hay que tener en cuenta que, de acuerdo a la definición de Neches, una ontología incluye no sólo los términos que están definidos explícitamente en ella, sino también el conocimiento que puede ser inferido de ella.

Unos años más tarde, en 1993, Gruber definió una ontología de la siguiente forma (Gruber, 1993):

“*Una ontología es una especificación explícita de una conceptualización*”.

Esta definición se convirtió en la más citada en la literatura y por la comunidad ontológica.

Basándose en la definición de Gruber, surgieron otras muchas definiciones de ontología. En 1997, Borst modificó ligeramente la definición de Gruber proponiendo lo siguiente (Borst, 1997):

“*Las ontologías se definen como una especificación formal de una conceptualización compartida*”.

Las definiciones de Gruber y Borst fueron fusionadas y explicadas por Studer y colegas como sigue (Studer, et al., 1998):

“Una ontología es una especificación formal y explícita de una conceptualización compartida. Por conceptualización se entiende un modelo abstracto de algún fenómeno del mundo, que identifica los conceptos relevantes de ese fenómeno. Explícita significa que el tipo de conceptos y las restricciones utilizadas han sido definidos explícitamente. Formal se refiere al hecho de que la ontología debería ser legible por la computadora. Compartida refleja la noción de que una ontología captura conocimiento consensuado, es decir, no es un conocimiento privado de algún individuo, sino aceptado por un grupo”.

Finalmente, se puede realizar una definición extensiva de ontología, indicando los componentes que la forman. En general, las ontologías proporcionan un vocabulario común de un área y definen, a diferentes niveles de formalismo, el significado de los términos y relaciones entre ellos. El conocimiento en ontologías se formaliza principalmente usando seis tipos de componentes: clases, atributos, relaciones, funciones, axiomas e instancias (Gruber, 1993):

1. Una **clase** puede ser una entidad sobre la que se dice algo, como por ejemplo un tipo de objeto, la descripción de una tarea, función, acción, estrategia, proceso de razonamiento, etc. Se suelen usar indistintamente los términos “clase” y “concepto”.
2. Los **atributos** representan la estructura interna de los conceptos. Atendiendo a su origen, los atributos se clasifican en **específicos** y **heredados**. Los atributos específicos son aquellos que son propios del concepto al que pertenecen, mientras que los heredados vienen dados por las relaciones taxonómicas en las que el concepto desempeña el rol de hijo y, por tanto, hereda los atributos del padre. Los atributos vienen caracterizados por el dominio en el cual pueden tomar valor.
3. Las **relaciones** representan un tipo de interacción entre los conceptos del dominio. Se definen formalmente como cualquier subconjunto de un producto cartesiano de n conjuntos, esto es: $R: C_1 \times C_2 \times \dots \times C_n$. Entre los distintos tipos de relaciones posibles, se encuentran las relaciones **taxonómicas** (“es_un”) y las **mereológicas** o **partonómicas** (“parte_de”) como relaciones binarias más destacadas.

4. Las **funciones** son un tipo especial de relaciones en las que el n-ésimo elemento de la relación es único para los n-1 precedentes. Formalmente, se definen las funciones (F) como: $F: C_1 \times C_2 \times \dots \times C_{n-1} \rightarrow C_n$. Como ejemplos, se pueden mencionar las funciones “madre de” y “precio de un coche usado”.
5. Los **axiomas** son expresiones que siempre son ciertas. Pueden ser incluidas en una ontología con muchos propósitos, tales como definir el significado de los componentes ontológicos, definir restricciones complejas sobre los valores de los atributos, argumentos de relaciones, etc., verificando la corrección de la información especificada en la ontología o deduciendo nueva información.
6. Por último, las **instancias** son las ocurrencias en el mundo real de los conceptos. En una instancia todos los atributos del concepto tienen asignado un valor concreto.

Las clases en la ontología se suelen organizar en taxonomías. En todo caso, cabe destacar que ontología y taxonomía son dos elementos diferentes, aunque algunas veces la noción de ontología se diluye en el sentido que las taxonomías se consideran ontologías completas (Studer, et al., 1998). Por ejemplo, UNSPSC, e-cl@ass, y RosettaNet, propuestas de estándares en el dominio del comercio electrónico (*e-Commerce*) y el Directorio Yahoo, una taxonomía para buscar en la Web, se consideran ontologías porque proporcionan una conceptualización consensuada de un dominio dado (Lassila & McGuinness, 2001). Al respecto, la comunidad de trabajo sobre ontologías distingue ontologías que son principalmente taxonomías de ontologías que modelan el dominio en una forma más profunda y proporcionan mayores restricciones en la semántica del dominio. La comunidad les da el nombre de *lightweight* y *heavyweight ontologies* respectivamente (ontologías ligeras y ontologías pesadas). Por un lado, las ontologías ligeras incluyen conceptos, taxonomías de conceptos, relaciones entre conceptos y propiedades que describen conceptos. Por otro lado, las ontologías pesadas añaden axiomas y restricciones a las ontologías ligeras. Los axiomas y las restricciones clarifican el significado de los términos agrupados en la ontología.

Como conclusión, se puede decir que las ontologías persiguen capturar conocimiento consensuado de forma genérica, y que pueden reutilizarse y compartirse

entre aplicaciones software y grupos de personas. Normalmente se construyen de forma cooperativa por diferentes grupos de personas en diferentes lugares (Gómez-Pérez, 1994).

2.1.2 Tipos de ontologías

Varios autores han tratado de categorizar los diferentes tipos de ontologías existentes, atendiendo a diferentes criterios. En este apartado, se presentan las clasificaciones de ontologías más relevantes.

En 1998, Guarino (Guarino, 1998) propone una clasificación basada en la generalidad de las ontologías. Clasifica las ontologías en:

- **Ontologías de alto nivel (*upper-level*)**. Ontologías que describen conceptos genéricos como tiempo, espacio y eventos. Son, en principio, independientes del dominio y se pueden reutilizar para construir nuevas ontologías.
- **Ontologías del dominio**. Ontologías que describen el vocabulario de un dominio dado, mediante la especialización de conceptos proporcionados por ontologías de alto nivel.
- **Ontologías de tarea**. Ontologías que describen el vocabulario requerido para llevar a cabo tareas o actividades genéricas, de nuevo especializando conceptos de ontologías de alto nivel.
- **Ontologías de aplicación**. Ontologías que describen el vocabulario de una aplicación específica, correspondiente, en general, a los roles desempeñados por entidades en un dominio dado durante la realización de alguna tarea o actividad.

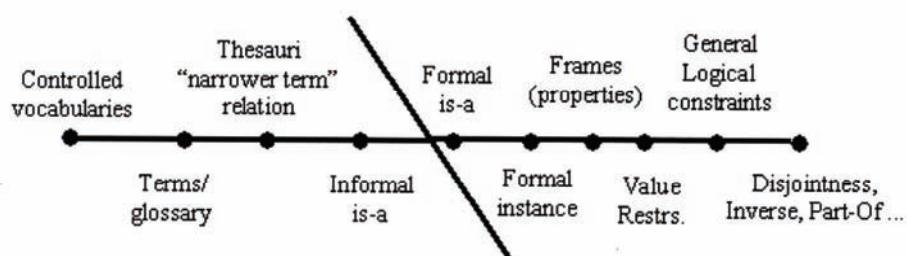


Figura 2.2. Categorización de ontologías de Lassila y McGuinness (Lassila & McGuinness, 2001).

Otra clasificación es la de Lassila y McGuinness (Lassila & McGuinness, 2001), que categoriza las ontologías en base a su estructura interna y contenido, siguiendo una línea donde las ontologías van desde “ligeras” a “pesadas”, dependiendo de la complejidad y sofisticación de los elementos que contienen. Esta clasificación es la siguiente (ver figura 2.2):

- **Vocabularios controlados.** Es decir, una lista finita de términos. Un ejemplo típico de esta categoría es un catálogo.
- **Glosarios.** Son listas de términos con su significado especificado en lenguaje natural. El formato de un glosario es similar al de un diccionario. Los términos se organizan en orden alfabético, acompañados por sus definiciones.
- **Tesauros.** Listas de términos y definiciones que estandarizan palabras para propósitos de indexado. Además de las definiciones, un tesauro también proporciona relaciones jerárquicas, asociativas y de equivalencia (i.e. sinónimos) entre términos.
- **Jerarquías informales is-a.** Son jerarquías que utilizan la relación de generalización de manera informal, es decir, conceptos relacionados pueden encontrarse agregados en una categoría incluso si no respectan la relación de generalización de forma estricta. Por ejemplo, los términos *alquiler de coches* y *hotel* no son tipos de viajes, pero pueden modelarse en una jerarquía informal como “subclases” del concepto *viaje* por tratarse de aspectos importantes del mismo.
- **Jerarquías formales is-a.** Utilizan la relación de generalización de forma estricta, mediante el concepto formal de “subclase”. Por ejemplo, las subclases del concepto *viaje* podrían ser: *vuelo*, *viaje en tren*, etc., pero no *alquiler de coches*.
- **Jerarquías formales is-a que incluyen instancias del dominio.** En este caso, se permite incluir instancias de los conceptos. Por ejemplo, una instancia de *vuelo* sería: *el vuelo AA7462 aterriza en Seattle, parte el 8 de febrero y cuesta 300\$*.
- **Marcos (frames).** La ontología incluye clases y sus propiedades, que pueden ser heredadas por clases de niveles inferiores en la taxonomía. Por ejemplo, un

vuelo tiene una *fecha de salida* y una *fecha de llegada*, un *nombre de compañía* y un *precio*. Todos estos atributos son heredados por las subclases del concepto *vuelo*.

- **Ontologías que expresan restricciones de valor.** En estas ontologías es posible establecer restricciones referentes a los valores que puede tomar una propiedad. Por ejemplo, el tipo de la propiedad *fecha de llegada* es *date*.
- **Ontologías que expresan restricciones lógicas generales.** Son las que presentan más expresividad, ya que en ellas es posible especificar restricciones lógicas entre términos usando lenguajes de definición de ontologías. Una restricción lógica en el contexto de un *viaje* es: *no es posible viajar desde USA a Europa en tren*.

Finalmente, Gómez-Pérez y colegas (Gómez-Pérez et al., 2004) proponen una clasificación que utiliza el tipo de información representado en la ontología como principal criterio. Clasifican las ontologías en:

- **Ontologías para la representación de conocimiento.** Ofrecen los elementos utilizados en representaciones basadas en *frames*, como clases, subclases, valores, atributos y axiomas.
- **Ontologías generales o comunes.** Representan conocimiento de sentido común, reutilizable entre diferentes dominios. Incluyen vocabulario relacionado con cosas, eventos, tiempo, espacio, causalidad y comportamiento, etc.
- **Ontologías de alto nivel (*top level* o *upper level*).** Describen conceptos generales.
- **Ontologías del dominio.** Ofrecen conceptos reutilizables en un mismo dominio.
- **Ontologías de tarea.** Describen el vocabulario relacionado con una tarea o actividad.
- **Ontologías de dominio-tarea.** Son ontologías de tarea que pueden ser reutilizadas en un dominio específico, pero no (en general) en dominios similares.
- **Ontologías de método.** Proporcionan definiciones para conceptos y relaciones relevantes en un proceso determinado.

- **Ontologías de aplicación.** Contienen todos los conceptos necesarios para modelar la aplicación en cuestión. Se utilizan para especializar y extender ontologías de dominio o de tarea para aplicaciones concretas.

2.1.3 Lenguajes de representación de ontologías

Durante los últimos años, se han desarrollado un conjunto de lenguajes que permiten representar las ontologías formalmente, de tal manera que éstas puedan ser procesadas automáticamente. En este apartado, se explica brevemente cómo han ido evolucionando estos lenguajes, desde su aparición hasta la actualidad.

A principios de los 90, se creó un conjunto de lenguajes de ontologías basados en Inteligencia Artificial. Básicamente, los paradigmas de representación de conocimiento (*Knowledge Representation, KR*) subyacentes a tales lenguajes estaban basados en lógica de primer orden (e.g., KIF), en marcos combinados con lógica de primer orden (e.g., CycL, Ontolingua, OCML y FLogic), y en lógica descriptiva (e.g., LOOM). OKBC se creó como un protocolo para acceder a ontologías implementadas en diferentes lenguajes con un paradigma de KR basado en marcos. La organización completa de estos lenguajes se muestra en la figura 2.3.

El primer lenguaje de etiquetado de ontologías en aparecer fue **SHOE** (Luke & Heflin, 2000). SHOE es un lenguaje que combina marcos y reglas. Se construyó como una extensión de HTML, en 1996. Se usaron etiquetas (*tags*) diferentes de los de la especificación HTML para permitir la inserción de ontologías en documentos HTML. Más tarde su sintaxis se adaptó a XML.

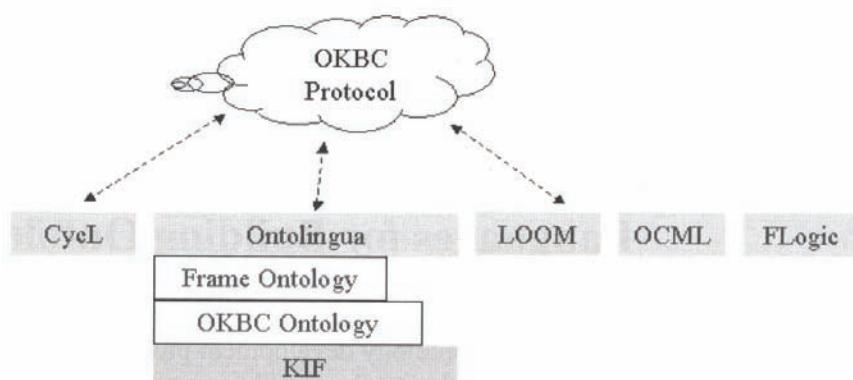


Figura 2.3. Lenguajes de representación de ontologías tradicionales (Gómez-Pérez, et al., 2004).

El resto de los lenguajes de etiquetado de ontologías presentados aquí están basados en XML. **XOL** (Karp et al., 1999) se desarrolló como una conversión a XML de un pequeño subconjunto de primitivas del protocolo OKBC, llamado OKBC-Lite. **RDF** (Lassila & Swick, 1999) fue desarrollado por el W3C (World Wide Web Consortium) como un lenguaje basado en una red semántica para describir recursos Web. Su desarrollo comenzó en 1997 y fue propuesto como una recomendación del W3C en 1999. El lenguaje **RDF Schema** (Brickley & Guha, 2003) fue construido por el W3C como una extensión al RDF con primitivas basadas en marcos. Este lenguaje se propuso como una recomendación candidata del W3C en el 2000. En noviembre del 2002 se revisó y se publicó su documento de referencia como un Borrador de Trabajo del W3C (W3C Working Draft). Posteriormente fue revisado de nuevo, en enero del 2003. La combinación de ambos lenguajes, RDF y RDF Schema, se conoce como **RDF(S)**.

Estos lenguajes han establecido los fundamentos de la Web Semántica. En este contexto, se desarrollaron tres lenguajes adicionales, como extensiones a RDF(S): OIL, DAML+OIL, y OWL. **OIL** (Fensel et al., 2000) se desarrolló a comienzos del 2000 en el marco del proyecto europeo de Tecnologías de la Sociedad de la Información On-To-Knowledge. Añade primitivas de KR basadas en marcos a RDF(S) y su semántica formal está basada en lógica descriptiva. **DAML+OIL** (Horrocks & van Harmelen, 2000) se creó más tarde (entre los años 2000 y 2001) por medio de un comité mixto de los Estados Unidos y de la Unión Europea en el contexto del proyecto de DARPA DAML. Se basó en la especificación previa DAML-ONT, que fue construida a finales del 2000, y en OIL. DAML+OIL añade primitivas de KR basadas en lógica descriptiva a RDF(S). En 2001 el W3C formó un grupo de trabajo llamado Web-Ontology (WebOnt) Working Group. El objetivo de este grupo era desarrollar un nuevo lenguaje de etiquetado de ontologías para la Web Semántica. El resultado de su trabajo es el lenguaje **OWL** (Ontology Web Language) (McGuinness & Van Harmelen, 2004), que es una recomendación del W3C desde el 10 de febrero de 2004. OWL cubre la mayor parte de las características de DAML+OIL y ha renombrado la mayoría de las primitivas que aparecían en ese lenguaje.

OWL se divide en tres niveles, dialectos o *species*: OWL Full, OWL DL y OWL Lite (ordenados de mayor a menor expresividad). La existencia de estas tres variantes del

lenguaje se debe a la necesidad de proporcionar soporte a diferentes tipos de usuarios, que necesitan distintos niveles de capacidad expresiva. Se puede decir que **OWL Full** es el lenguaje OWL al completo, pues cuenta con todas las construcciones del lenguaje OWL y, por tanto, proporciona una expresividad máxima. RDF es un subconjunto de este lenguaje. Una alternativa más eficiente computacionalmente a OWL Full es **OWL DL**. El principal objetivo de este sublenguaje es proporcionar un dialecto que soporte aplicaciones de razonamiento. OWL DL utiliza las mismas construcciones que OWL Full, pero con algunas restricciones en su utilización. Por último, **OWL Lite** es el lenguaje más simple de los tres, pues soporta únicamente un subconjunto de las construcciones de OWL Full, algunas de ellas de forma restringida. OWL Lite está orientado a aquellos usuarios que quieren beneficiarse de las ventajas derivadas de utilizar ontologías sin necesidad de codificar relaciones semánticas complejas. Proporciona el soporte suficiente para actualizar bases de datos e información en XML o RDF(S) a representaciones de información basadas en ontologías en OWL. La figura 2.4 muestra las relaciones existentes entre estos dialectos. Al igual que DAML+OIL, OWL se construye sobre RDF(S). Por lo tanto, algunas primitivas de RDF(S) son reutilizadas por OWL, y las ontologías en OWL se escriben o bien en XML o bien con la notación de RDF.

Además de los lenguajes de representación de ontologías de propósito general como OWL o RDF, también se han desarrollado lenguajes para dominios específicos. En el dominio de la biomedicina, el consorcio Gene Ontology (GO consortium)³ ha ideado un formato que supone una alternativa a formatos estándar de propósito general. Este formato se conoce como el formato OBO (OBO format)⁴ y ha sido diseñado para proporcionar alta legibilidad a las personas y facilitar el *parsing* a las computadoras, además de simplificar la extensibilidad con una redundancia mínima. Actualmente, el formato OBO es un formato ampliamente utilizado para representar ontologías biomédicas y existen diversas utilidades software para trabajar con él. Un ejemplo es el *plugin* para Protégé conocido como OBO Converter (Moreira & Musen, 2007), que permite convertir ontologías de formato OBO a OWL y viceversa. En la

³ <http://www.geneontology.org/GO.consortiumlist.shtml/>

⁴ http://www.geneontology.org/GO.format.obo-1_2.shtml/

figura 2.5 se puede ver un fragmento del fichero de representación de la ontología African Traditional Medicine⁵ en formato OBO.

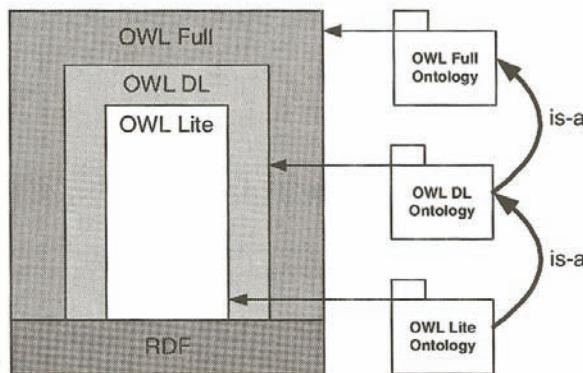


Figura 2.4. Relaciones existentes entre los distintos dialectos de OWL (Lacy, 2005).

```

<...>

[Term]
id: ATM:00062
name: potion to drink
namespace: african_traditional_practices
is_a: ATM:00053 ! potion

[Term]
id: ATM:00063
name: potion to mass
namespace: african_traditional_practices
is_a: ATM:00053 ! potion

[Term]
id: ATM:00064
name: potion for specific dose
namespace: african_traditional_practices
is_a: ATM:00053 ! potion

[Term]
id: ATM:00065
name: potion to anoint
namespace: african_traditional_practices
is_a: ATM:00053 ! potion

[Term]
id: ATM:00066
name: potion to purge
namespace: african_traditional_practices
is_a: ATM:00053 ! potion

<...>

```

Figura 2.5. Fragmento de la ontología African Traditional Medicine en formato OBO.

⁵ <http://bioportal.bioontology.org/ontologies/40223/>

2.1.4 Aplicaciones de las ontologías

El continuo desarrollo y aplicación de tecnologías y técnicas computacionales a grandes volúmenes de información ha provocado que a día de hoy existan múltiples dominios (e.g. medicina y biomedicina, industria, comercio, finanzas, etc.) que dependen de grandes cantidades de información que se encuentra almacenada utilizando terminologías y formatos diferentes, lo cual constituye una importante barrera a la hora de procesarla y reutilizarla de manera integrada. Debido a esto, el desarrollo de métodos y tecnologías que permitan organizar, compartir y acceder a esta información se ha convertido en una prioridad en los últimos años.

Las descripciones en texto libre son una forma fácil y cómoda de mantener grandes niveles de expresividad, pero los resultados no son fácilmente entendibles y reutilizables por otras personas. Además, este tipo de descripciones son difíciles de procesar usando una computadora. Como una alternativa al texto libre, las ontologías se consideran una forma práctica de describir todas las entidades en un dominio y las relaciones entre ellas de forma estandarizada, facilitando el intercambio de información entre personas o sistemas que utilizan diferentes representaciones para el mismo conocimiento. Y lo que es incluso más importante, el conocimiento representado utilizando ontologías puede ser procesado por las computadoras, permitiendo llevar a cabo búsquedas, análisis, inferencias y minería de datos a gran escala y de forma automática (Ashburner et al., 2000; Gruber, 1995).

En la actualidad, las ontologías se están utilizando en diversas aplicaciones pertenecientes a diferentes áreas. Así, por ejemplo, existen múltiples aplicaciones de las ontologías en salud, medicina y biomedicina (e.g. (Rubin et al., 2007), (Smith & Brochhausen, 2008)), aprendizaje (e.g. (Gaeta et al., 2009)), administración y gobierno (e.g. (Grandi et al., 2009)), finanzas (e.g. (Zhang et al., 2000)), turismo (e.g. (Barta et al., 2009)), sistemas de información geográfica (e.g. (Rotondo, 2010)), industria (e.g. (Borgo & Lesmo, 2008)), comercio (e.g. (Liu et al., 2009)), etc.

Además de estas aplicaciones, las ontologías se consideran un instrumento básico en otra importante iniciativa, que a su vez hace posible el funcionamiento y soporte de la mayoría de las aplicaciones previamente mencionadas. Esta iniciativa se conoce con el nombre de Web Semántica (o Semantic Web) y pretende mejorar la actual World

Wide Web, ampliando la interoperabilidad entre los sistemas informáticos y reduciendo la necesaria mediación de operadores humanos. Este objetivo se pretende conseguir mediante la inclusión de metadatos semánticos en la actual Web (en base a ontologías), que describan los actuales contenidos, haciendo posible su procesamiento automático por parte de las computadoras. Las ontologías se consideran el instrumento básico en la Web Semántica, ya que son una buena forma de especificar conceptos de manera formal, explícita y consensuada, y su aplicación a la Web Semántica ha marcado una nueva etapa en su evolución.

Debido a su relevancia, el concepto de Web Semántica se tratará en detalle más adelante en este capítulo. En el siguiente apartado, se trata el tipo de ontologías que resultan de más interés en el contexto de esta tesis, las ontologías en campos “bio” o “bio-ontologías”.

2.1.5 Bio-ontologías

En la era actual, posterior a la secuenciación del genoma, la aplicación de tecnologías y técnicas computacionales de alto rendimiento a grandes volúmenes de datos genómicos, fisiológicos y de enfermedades, está dando como resultado un crecimiento exponencial de información sobre diferentes moléculas y factores. Manejar esta información adecuadamente es un aspecto crucial para entender los mecanismos moleculares de enfermedades multifactoriales, como el cáncer, y realizar así avances significativos en áreas como el desarrollo de fármacos o las aproximaciones de diagnóstico. El potencial de las ontologías de combinar altos niveles de expresividad con todas las ventajas de la computación explica por qué las ontologías están ganando cada vez más popularidad en lo que respecta a la representación del conocimiento en campos “bio” (Martínez-Romero et al., 2010).

En “bio-dominios”, como la biomedicina o la bioquímica, las ontologías están formadas por conceptos biológicos y relaciones entre ellos. Su aplicación más común es la anotación basada en ontologías de datos básicos, o lo que es lo mismo, la asociación de elementos de ontologías (i.e. conceptos y relaciones) a datos, de tal manera que las ontologías puedan ser usadas por personas y computadoras para compartir, buscar y navegar a través de información genética, fenotípica y de enfermedades. Estas ontologías, conocidas comúnmente como “bio-ontologías”

(*bio-ontologies*) (Blake, 2004), se están utilizando cada vez más en diversas aplicaciones bioinformáticas, que varían desde la anotación y búsqueda semánticas al análisis de datos a gran escala. Para representar este tipo de ontologías, el consorcio Gene Ontology (GO consortium)⁶ ha ideado un formato específico, el formato OBO, ya mencionado en el apartado 2.1.3.

2.1.5.1 Ejemplos de bio-ontologías

En este apartado, se muestran algunos ejemplos de las “bio-ontologías” más populares de la actualidad.

2.1.5.1.1 La Clasificación Internacional de Enfermedades

La Clasificación Internacional de Enfermedades (CIE)⁷ es la clasificación de diagnósticos estándar de la Organización Mundial de la Salud (OMS) para propósitos epidemiológicos, de gestión de la salud y de uso clínico. Esto incluye el análisis de la situación sanitaria general de grupos de la población y la monitorización de la incidencia y prevalencia de enfermedades y otros problemas de salud en relación con otras variables como las características y las circunstancias de los individuos afectados, localización de recursos, aspectos de calidad, etc.

Esta ontología se utiliza para clasificar enfermedades y otros problemas de salud almacenados en diversos tipos de registros de salud y vitales, incluyendo certificados de defunción y registros sanitarios. Además de permitir el almacenamiento y la recuperación de información diagnóstica para propósitos epidemiológicos y clínicos, estos registros también proporcionan la base para la recopilación de estadísticas de mortalidad y morbilidad a nivel nacional para los estados miembros de la OMS. En la figura 2.6 se puede ver, de forma gráfica, un fragmento de esta ontología.

La versión más reciente de la CIE es la CIE-10, que comenzó a utilizarse en los estados pertenecientes a la OMS en 1994. Esta clasificación es la más reciente de una serie originada en la década de 1850. La primera edición, conocida como la “Lista Internacional de Causas de Mortalidad”, fue adoptada por el Instituto Internacional de Estadística en 1893. En el momento de la creación de la OMS, en 1948, ésta acepta

⁶ <http://www.geneontology.org/GO.consortiumlist.shtml/>

⁷ <http://www.who.int/classifications/icd/en/>

toda la responsabilidad sobre la CIE, cuya sexta edición acababa de publicarse. Esta sexta edición era la primera que incluía causas de morbilidad. En el año 1967 se estipula el uso de la CIE para las estadísticas de mortalidad y morbilidad en todos los países miembros de la OMS.

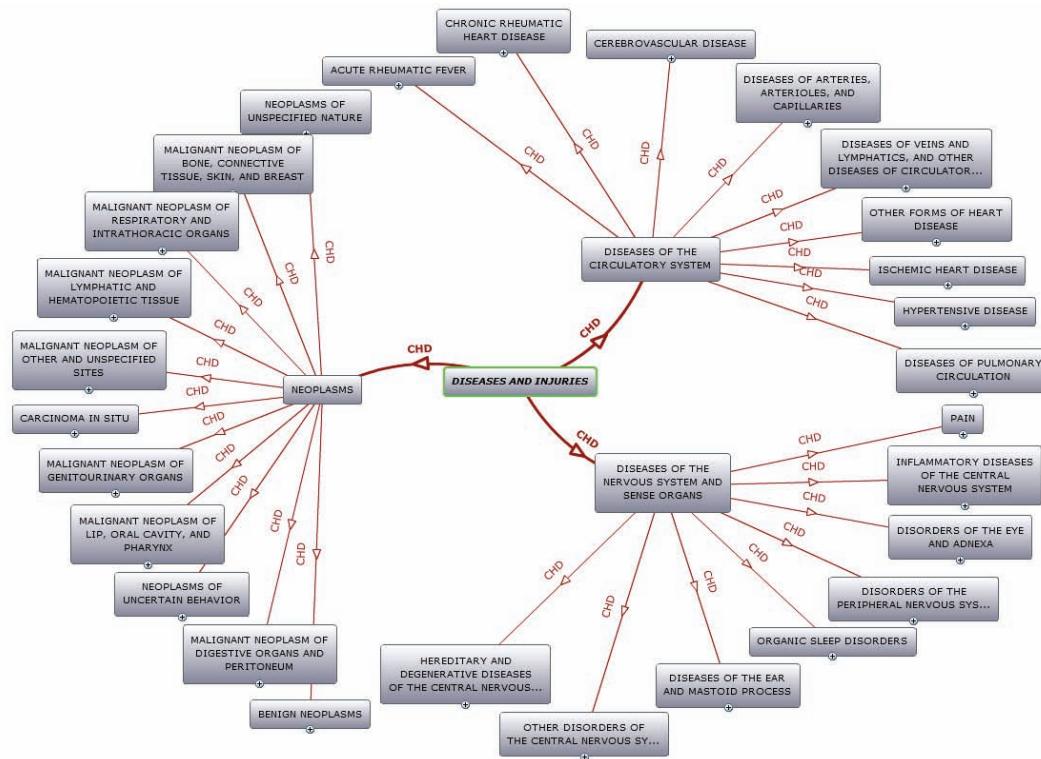


Figura 2.6. Categorías de la ontología CIE relativas a los ámbitos de Neurología, Cardiología y Oncología. La relación “CHD” es la relación “subclase de” (o is-a). Generada mediante BioPortal.

La clasificación CIE se puede consultar a través del sitio Web de la OMS⁸. También se puede descargar en formato OWL desde el sitio Web BioPortal, del Centro Nacional para la Ontología Biomédica de EE.UU. (NCBO)⁹.

2.1.5.1.2 El NCI Thesaurus

El NCI Thesaurus es una amplia ontología que abarca el dominio del cáncer, incluyendo enfermedades, descubrimientos y anomalías relacionadas con el cáncer; anatomía; agentes, medicamentos y productos químicos; genes, etc. Es una ontología que ha sido creada por el Instituto Nacional del Cáncer de EE.UU. (National Cancer

⁸ <http://apps.who.int/classifications/apps/icd/icd10online/>

⁹ <http://bioportal.bioontology.org/ontologies/35686>

Institute, NCI), el cual a día de hoy se encarga de su mantenimiento. En ciertas áreas, como las enfermedades relacionadas con el cáncer y las quimioterapias de combinación, proporciona la terminología más granular y consistente que existe. Combina terminología procedente de numerosos dominios relacionados con la investigación del cáncer, y proporciona una forma de integrar o enlazar toda esta información mediante relaciones semánticas. En la figura 2.7 se muestra un fragmento de esta ontología.

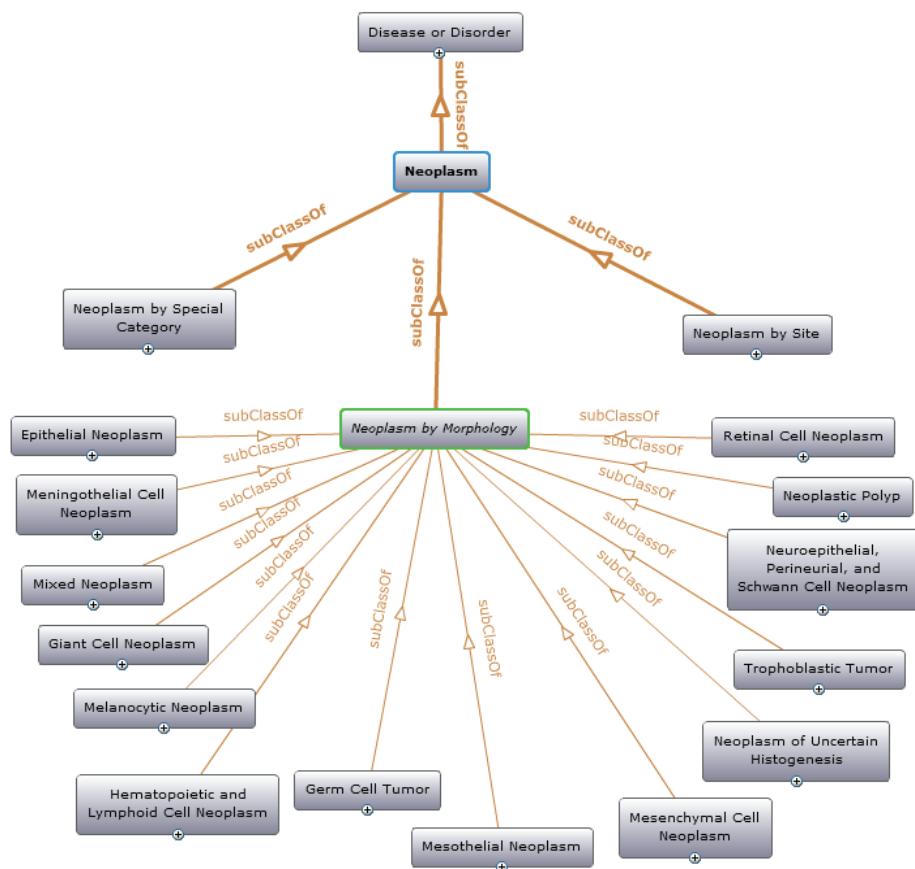


Figura 2.7. Fragmento de la ontología NCI Thesaurus, en la que se muestran algunos conceptos relacionados con los tipos de neoplasias, detallando la clasificación de los neoplasmas de acuerdo a su morfología.

Actualmente, el NCI Thesaurus contiene más de 34.000 conceptos, estructurados en 20 árboles taxonómicos. Además, también almacena los cambios producidos en los conceptos que conforman la ontología a lo largo del tiempo, a medida que el conocimiento científico evoluciona. Dentro del NCI, esta ontología se utiliza para

proporcionar soporte terminológico al portal Web del organismo¹⁰, a otros múltiples portales que dan soporte a consorcios y otros organismos de investigación, y también como la base semántica sobre la cual se construyen los objetos de los portales del NCICB (NCI Center for Bioinformatics). Esta ontología se encuentra públicamente accesible a través de una licencia gratuita en diversos formatos, incluyendo el lenguaje de representación de ontologías OWL.

2.1.5.1.3 SNOMED CT

SNOMED CT (Systematized Nomenclature of Medicine-Clinical Terms) es una extensa ontología sobre el cuidado clínico de la salud, ampliamente consensuada científicamente, que en la actualidad se utiliza en más de 50 países y que se puede considerar fundamental para los registros de salud electrónicos. SNOMED CT fue creada originalmente por la Escuela de Patólogos de América (College of American Pathologists), a partir de SNOMED RT y una clasificación conocida como Clinical Terms Version 3, elaborada por el Departamento de Salud del Reino Unido.

El principal objetivo de SNOMED CT es mejorar el cuidado del paciente mediante el desarrollo de sistemas que permitan registrar de forma precisa los eventos relacionados con su salud. Para ello, la ontología proporciona la terminología general que representa el núcleo de los registros de salud electrónicos (*Electronic Health Records*, EHR) y contiene más de 311.000 conceptos activos con significado único y definiciones lógicas, basadas en jerarquías. Cuando se implementa en aplicaciones software, SNOMED CT se puede utilizar para representar de forma consistente, fiable y exhaustiva, información relevante desde un punto de vista clínico, como parte integral de la producción de registros de salud electrónicos. En la figura 2.8 se puede observar el crecimiento de SNOMED CT a través del número de códigos de concepto de esta ontología activos durante los últimos seis años.

Para utilizar SNOMED CT se necesita licencia. En el sitio Web de la IHTSDO (International Health Terminology Standards Development Organisation)¹¹ se puede encontrar información sobre los pasos a realizar para obtener una licencia.

¹⁰ <http://cancer.gov>

¹¹ <http://www.ihtsdo.org/>

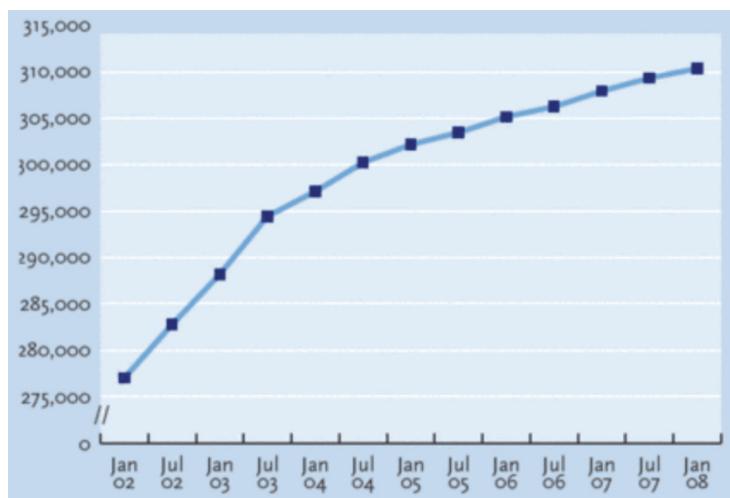


Figura 2.8. Códigos de concepto de SNOMED CT activos durante el periodo Enero 02 – Enero 08 (extraída del sitio Web de SNOMED CT¹²).

2.1.5.1.4 El proyecto Gene Ontology

El proyecto Gene Ontology (GO) es un esfuerzo de colaboración dirigido a proporcionar descripciones consistentes de genes y productos génicos de organismos. El proyecto comenzó en 1998 como una colaboración entre tres bases de datos sobre organismos diferentes: FlyBase¹³ (*Drosophila*), la Saccharomyces Genome Database¹⁴ (SGD) y la Mouse Genome Database¹⁵ (MGD). Desde entonces, el Consorcio GO ha crecido para incluir muchas otras bases de datos, entre las que se encuentran algunos de los mayores repositorios sobre genomas de plantas, animales y microbios. En el sitio Web del Consorcio GO¹⁶ se puede ver una lista completa de las organizaciones que forman parte del mismo.

El proyecto GO ha desarrollado tres vocabularios controlados estructurados (ontologías) que describen productos génicos en términos de (1) sus procesos biológicos asociados, (2) sus componentes celulares y (3) sus funciones moleculares, de forma independiente de la especie de que se trate. Un producto genético puede estar asociado a o localizado en uno o más componentes celulares; se encuentra activo en uno o más procesos biológicos, durante los que realiza una o más funciones

¹² <http://www.ihtsdo.org/snomed-ct/snomed-ct0>

¹³ <http://flybase.bio.indiana.edu/>

¹⁴ <http://www.yeastgenome.org/>

¹⁵ <http://www.informatics.jax.org/>

¹⁶ <http://www.geneontology.org/GO.consortiumlist.shtml>

moleculares. Por ejemplo, el producto génico “citocromo c” puede ser descrito por el término de función molecular “actividad oxidoreductasa”, los términos de proceso biológico “fosforilación oxidativa” e “inducción de muerte celular”, y los términos de componente celular “matriz mitocondrial” y “membrana interna mitocondrial”.

En la figura 2.9 se puede observar un fragmento de Gene Ontology, en el que se muestra el conjunto de términos pertenecientes al proceso biológico “pigmentación”. Las relaciones entre nodos se representan mediante flechas coloreadas, cuyo tipo viene indicado por la letra que etiqueta cada relación.

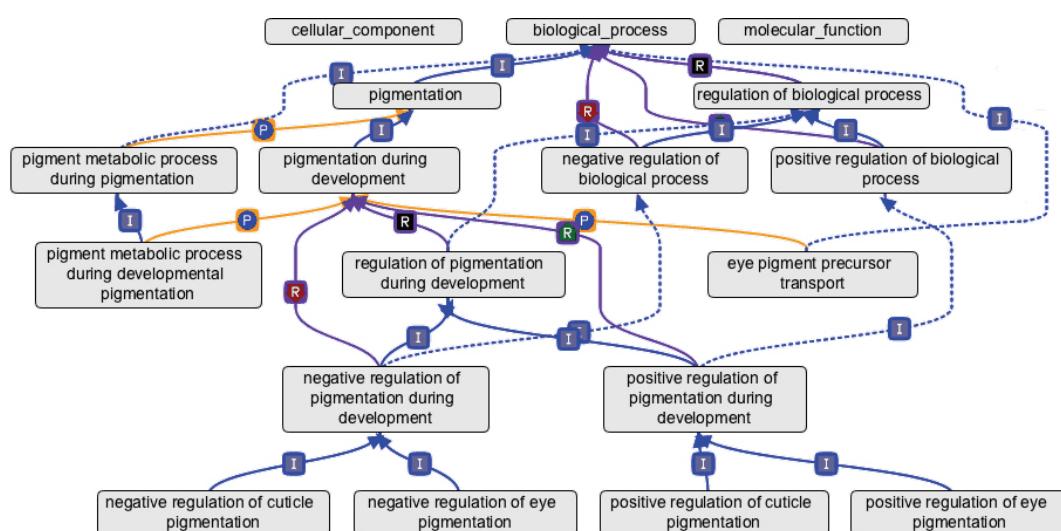


Figura 2.9. Conjunto de términos de Gene Ontology pertenecientes al proceso biológico “pigmentación” (extraída de www.geneontology.org).

En el proyecto GO, existen tres aspectos diferenciados: en primer lugar, el desarrollo y mantenimiento de las ontologías en sí mismas; en segundo lugar, la anotación de productos génicos, que conlleva realizar asociaciones entre las ontologías y los genes y productos génicos en las bases de datos con las que se colabore; y en tercer lugar, el desarrollo de herramientas que faciliten la creación, mantenimiento y uso de las ontologías.

Todos los datos del proyecto GO se encuentran disponibles de forma gratuita. Los datos de la ontología se pueden descargar en diferentes formatos (e.g. OWL, OBO, MySQL, etc.) desde el sitio Web de descargas del proyecto¹⁷.

¹⁷ <http://www.geneontology.org/GO.downloads.shtml>

2.1.5.1.5 Foundational Model of Anatomy

El Modelo Fundamental de Anatomía (Foundational Model of Anatomy, FMA), desarrollado y mantenido por el Grupo de Informática Estructural¹⁸ de la Universidad de Washington, es una fuente de conocimiento en continua evolución para los informáticos biomédicos, útil para el desarrollo de aplicaciones en campos como la educación, la medicina clínica, registros de salud electrónicos, la investigación biomédica y otras áreas de cuidado y gestión de la salud.

Se ocupa de la representación de clases o tipos y las relaciones necesarias para la representación simbólica de la estructura fenotípica del cuerpo humano, de forma que sea comprensible para los humanos además de navegable e interpretable por las máquinas. De forma más específica, el FMA es una ontología del dominio que representa un cuerpo coherente de conocimiento declarativo sobre la anatomía humana. Su base ontológica puede ser aplicada y extendida para otras especies.

En la actualidad, este modelo contiene aproximadamente 75.000 clases y 120.000 términos. Estas clases se encuentran relacionadas entre sí mediante más de 2,1 millones de relaciones, de 168 tipos diferentes. De acuerdo a estos datos, el modelo FMA se considera una de las mayores fuentes de conocimiento interpretable por máquinas en el ámbito de las ciencias biomédicas.

Esta ontología posee cuatro elementos relacionados entre sí:

- **Taxonomía de anatomía (At).** Clasifica entidades anatómicas de acuerdo a las características que comparten (*genus*) y diferencian (*differentia*) de otras entidades.
- **Abstracción estructural anatómica (ASA).** Especifica las relaciones parte-todo y espaciales que existen entre las entidades de At.
- **Abstracción de transformación anatómica (ATA).** Especifica la transformación morfológica de las entidades representadas en At durante el desarrollo prenatal y el ciclo de vida postnatal.

¹⁸ <http://sig.biostr.washington.edu/index.html>

- **Metaconocimiento (Mk).** Especifica los principios, reglas y definiciones de acuerdo a las que se representan las clases y las relaciones de los otros tres componentes de FMA.

El componente más extenso de FMA es la “Taxonomía de anatomía”, cuya clase principal es la “Estructura Anatómica” (*Anatomical Structure*). Las estructuras anatómicas incluyen todos los elementos generados por la expresión de grupos de genes de un organismo. Por ejemplo, incluyen macromoléculas biológicas, células y sus partes, porciones de tejidos, órganos y sus partes, así como sistemas de órganos y partes del cuerpo (regiones del cuerpo). En la figura 2.10 se presenta un fragmento de la ontología FMA en la que se puede ver la clasificación anatómica de tipos de órganos del cuerpo humano, mostrando los tipos de órganos “sólidos”.

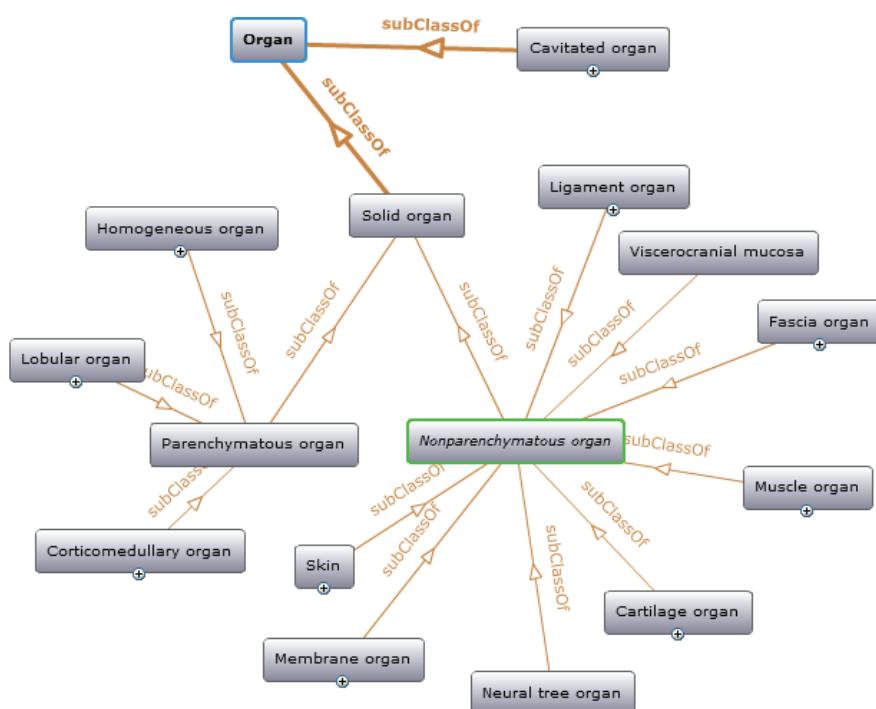


Figura 2.10. Fragmento de la ontología FMA.

2.1.5.2 Unified Medical Language System

El Unified Medical Language System (UMLS)¹⁹, es un conjunto de fuentes de conocimiento del dominio biomédico y herramientas que facilitan el acceso a dichas

¹⁹ <http://www.nlm.nih.gov/research/umls/>

fuentes, desarrollado por la National Library of Medicine de EE.UU. Está formado por tres elementos principales:

- Metatesauro (UMLS Metathesaurus)
- Red Semántica (UMLS Semantic Network)
- Herramientas léxicas (SPECIALIST Lexicon & Tools)

Estos tres elementos se pueden utilizar de forma independiente o conjunta. A continuación, se explica en qué consiste cada uno de ellos.

2.1.5.2.1 Metatesauro

El Metatesauro UMLS es un conjunto de ontologías que representan diferentes puntos de vista de la biomedicina, tanto en la práctica como en el ámbito de la investigación. Este recurso, integra más de 2 millones de términos, asociados a más de 1 millón de conceptos diferentes, de más de 100 familias de vocabularios biomédicos, así como 12 millones de relaciones entre estos conceptos. Todas las ontologías de las que se ha hablado en el apartado 2.1.5.1 están incluidas en UMLS.

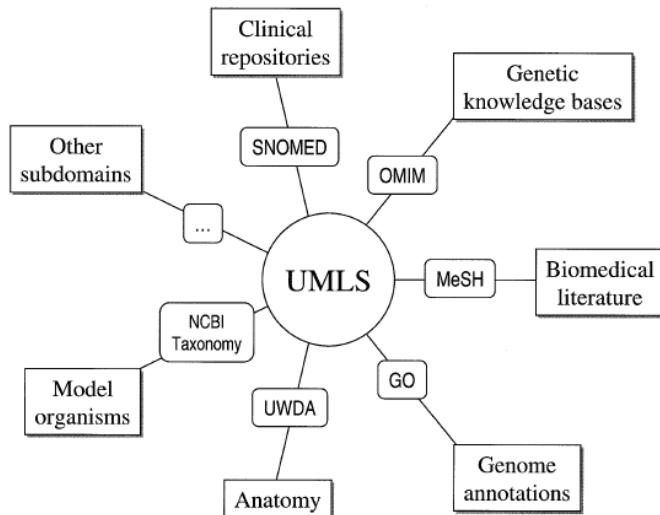


Figura 2.11. Subdominios incluidos en UMLS (Bodenreider, 2004).

En el diagrama de la figura 2.11 se muestra una visión general del Metatesauro, en el que se pueden ver algunas de las ontologías que contiene (e.g. SNOMED, OMIM, MeSH, etc.) y los dominios que éstas abarcan (e.g. Repositorios clínicos, Bases de datos genéticas, Literatura biomédica, etc.). El Metatesauro UMLS se encuentra disponible

en DVD y mediante transferencia FTP, de forma gratuita para propósitos de investigación y tras firmar un acuerdo de licencia. Viene acompañado de una herramienta software llamada MetamorphoSys que permite configurar e instalar una versión del Metatesauro para propósitos específicos. Este proceso de instalación permite almacenar en una base de datos toda la información (conceptos, nombres de concepto, relaciones, etc.) de los vocabularios que componen el Metatesauro (ver figura 2.12).

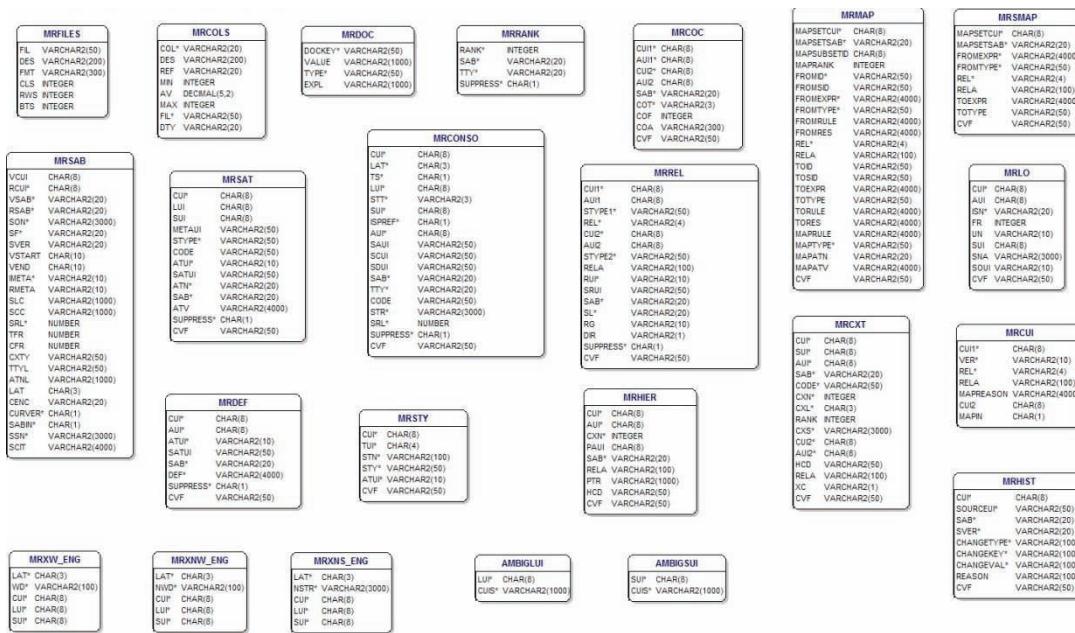


Figura 2.12. Diagrama de entidades del Metatesauro UMLS.

2.1.5.2.2 Red Semántica

La Red Semántica de UMLS (UMLS Semantic Network) es una ontología de categorías, o tipos semánticos (*semantic types*), cada una de las cuales agrupa un conjunto de conceptos del Metatesauro. Todos los conceptos en el Metatesauro disponen, como mínimo, de un tipo semántico. La Red Semántica proporciona una categorización consistente de todos los conceptos del Metatesauro y un importante conjunto de Relaciones Semánticas (*Semantic Relations*) entre tipos semánticos. En la figura 2.13 se muestra un fragmento de la Red Semántica, en la que se pueden diversos tipos semánticos y diferentes tipos de relaciones entre ellos.

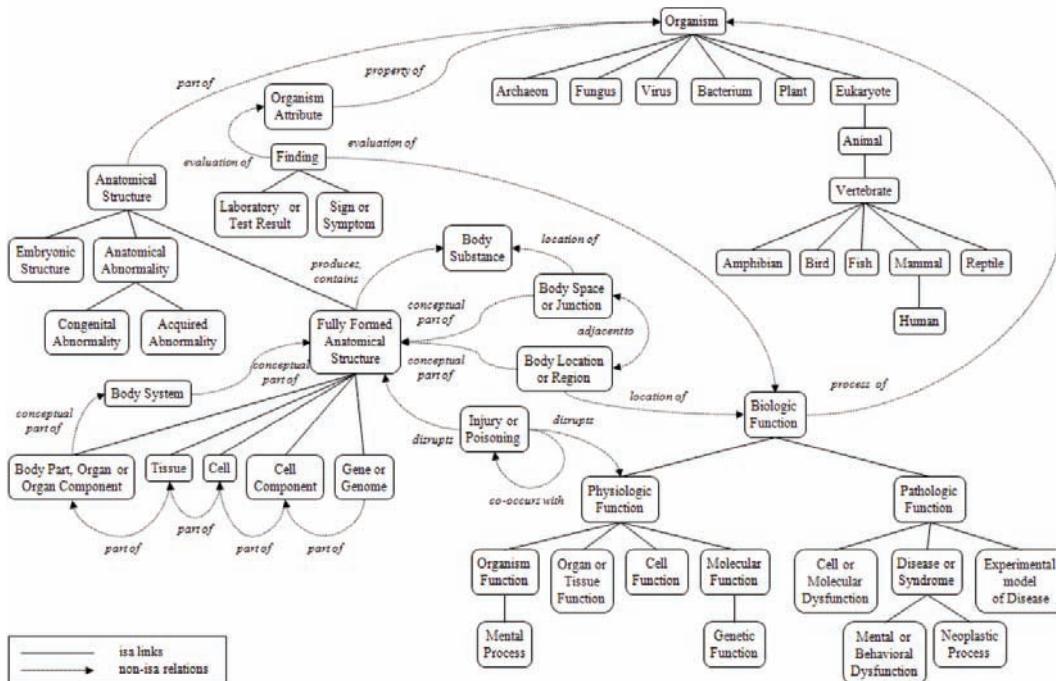


Figura 2.13. Fragmento de la Red Semántica de UMLS. Extraída del manual de referencia online de UMLS²⁰. Las líneas para las que no se muestra el tipo de relación se refieren a relaciones *is-a*.

A modo de ejemplo, el concepto *aorta* del Metatesauro, pertenece al tipo semántico *Body Part, Organ, or Organ Component*, mientras que el concepto *stomach disorder* está asociado al tipo semántico *Disease or Syndrome*. La Red Semántica se puede instalar como un conjunto de tablas de BD, que almacenan toda la información sobre los tipos semánticos y sus relaciones (ver figura 2.14).

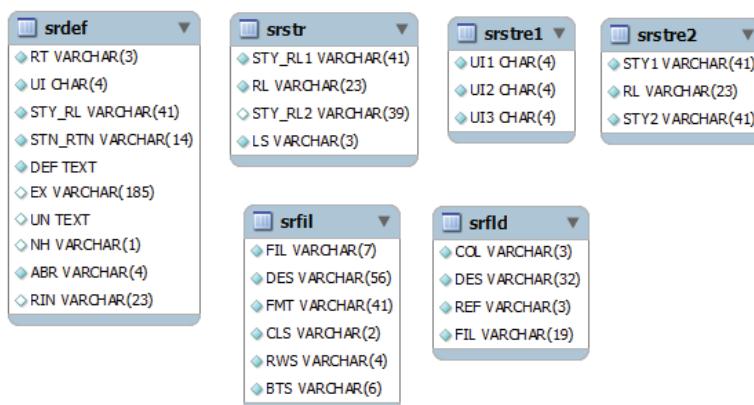


Figura 2.14. Tablas de BD de la Red Semántica.

²⁰ <http://www.ncbi.nlm.nih.gov/books/NBK9676/>

2.1.5.2.3 SPECIALIST Lexicon y Herramientas Léxicas

El SPECIALIST Lexicon es un léxico general del idioma inglés que incluye muchos términos biomédicos, tanto palabras comunes en inglés como vocabulario biomédico, y que ha sido desarrollado para su uso en tareas de procesado del lenguaje natural (*Natural Language Processing*, NLP). Para cada término o palabra en este recurso, se almacena toda la información sintáctica, morfológica y ortográfica necesaria requerida para llevar a cabo tareas de NLP.

Las Herramientas Léxicas (Lexical Tools) son un conjunto de utilidades diseñadas para afrontar el gran nivel de variabilidad de las palabras y términos en lenguaje natural. Habitualmente, las palabras tienen varias formas derivadas que se pueden considerar instancias de una misma palabra. Por ejemplo, el verbo “treat” tiene tres variantes: “treats” (tercera persona del singular del presente del verbo “treat”), “treated” (pasado y participio) y “treating” (gerundio). Además, los términos formados por múltiples palabras, como muchos de los pertenecientes al Metatesauro, pueden tener diversas formas dependientes del orden en el que se encuentran dichas palabras, además de las variantes básicas. El paquete de Herramientas Léxicas de UMLS está formado por tres utilidades: un normalizador, un generador de índices de palabras y un generador de variantes léxicas. Estas herramientas se explican brevemente a continuación. En el capítulo 6 del manual de referencia de UMLS²¹ se puede encontrar más información sobre su funcionamiento.

- **Normalizador (Normalizer o herramienta norm).** Esta herramienta recibe un término como entrada y crea un conjunto de términos “normalizados” con el mismo significado, que se utilizan en el índice de cadenas de texto normalizadas de UMLS (tabla MRXNS de la figura 2.12). Para usar el índice de cadenas de texto normalizadas, es necesario en primer lugar aplicar el normalizador. Esta utilidad transforma la cadena de texto original en una versión en minúsculas, sin puntuación (e.g. sin guiones), marcadores genitivos (como el apóstrofe que indica posesión en inglés), stop-words (e.g. *a*, *the*, *of*, etc.), acentos diacríticos (e.g. los existentes en Protégé), etc. También transforma verbos a infinitivo y nombres a singular. Para el caso de términos formados por

²¹ <http://www.ncbi.nlm.nih.gov/books/NBK9680/>

varias palabras, éstas se ordenan por orden alfabético. En algunas situaciones, la normalización no es única. Por ejemplo, el término *scleroses* puede referirse al plural del nombre *sclerosis* o bien a la tercera persona del singular del verbo *sclerose*. Por lo tanto, como resultado de la normalización de una lista de términos, la longitud de la lista normalmente aumenta, proporcionando términos adicionales que pueden resultar de utilidad.

- **Generador de índices de palabras (*Word Index Generator* o herramienta *wordInd*)**. Esta utilidad divide las cadenas de texto en palabras para su uso en el índice de palabras de UMLS (tabla MRXW de la figura 2.12).
- **Generador de variantes léxicas (*Lexical Variant Generator* o herramienta *Ivg*)**. Esta utilidad genera un conjunto de variantes léxicas para palabras de entrada. Consiste en varios componentes que se pueden combinar de varias formas para producir variantes léxicas. De hecho, el Normalizador de UMLS es básicamente la herramienta *hg* con una configuración específica.

2.1.6 Repositorios de ontologías

El gran número de ontologías existentes ha motivado el desarrollo de repositorios o librerías de ontologías que faciliten el acceso a ellas, para su reutilización. Algunos aspectos importantes de los repositorios de ontologías son los siguientes (Sabou, et al., 2006b):

- **Tamaño.** Uno de los aspectos que han cambiado de forma más radical con la evolución de la Web Semántica es el tamaño de los repositorios de ontologías. Los actuales repositorios, con capacidad para obtener ontologías de forma automática, son significativamente mayores que los repositorios tradicionales. Por ejemplo, el repositorio del buscador de ontologías Swoogle (Ding et al., 2004) contiene más de 10.000 ontologías, de diferentes dominios. Normalmente, estos repositorios disponen de alguna funcionalidad para realizar búsquedas sobre ellos (de nuevo, un claro ejemplo es el buscador Swoogle). Sin embargo, en general se trata de funcionalidades de búsqueda genéricas que no proporcionan buenos resultados. En la actualidad, es

prioritario disponer de mejores métodos para seleccionar las ontologías más adecuadas de estos repositorios.

- **Heterogeneidad, calidad y nivel de detalle.** Puesto que muchos de los actuales repositorios de ontologías disponen de métodos automáticos para rastrear la Web en busca de ontologías, la colección de ontologías resultante suele caracterizarse por una gran heterogeneidad, a diferentes niveles. Las ontologías recopiladas cubren un amplio rango de temáticas y existen diferencias importantes en cuanto a su calidad, que varía desde ontologías altamente formales, que han sufrido un riguroso proceso de revisión, hasta simples taxonomías creadas individualmente y no consensuadas. Además, el nivel de detalle de las ontologías existentes para un mismo dominio suele diferir en gran medida. Estos aspectos suponen nuevos retos a la hora de seleccionar la ontología con el nivel de calidad y granularidad más adecuado.
- **Nivel de control.** Algunos repositorios de ontologías proporcionan un entorno controlado a través del cual los usuarios pueden hacer públicas sus ontologías o someterlas a revisión al resto de la comunidad. Un ejemplo de este caso es BioPortal (Whetzel et al., 2009). Otros repositorios disponen de mecanismos de búsqueda automáticos (e.g. Swoogle (Ding, et al., 2004), Watson (d'Aquin et al., 2007), OntoSelect (Buitelaar et al., 2004), etc.). En este último caso, es posible construir repositorios con un gran número de ontologías rápidamente, pero la calidad de las mismas puede verse afectada.
- **Modularidad.** Algunos repositorios de ontologías consideran la modularidad de ontologías (acceso a una ontología de forma modular) como un prerequisito para la reutilización.

En la figura 2.15 se puede ver el sitio Web en el que se listan las ontologías del sistema de búsqueda OntoSelect (Buitelaar, et al., 2004).



Browse Ontologies																												
Show ontologies starting with:																												
<all>	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	0-9	«rest»
Showing 1 to 30 of 1530																										<< < 1 2 3 4 5 6 7 8 9 10 > >>		
Title	Domain	Format	Language	Labels	Classes	Properties	Included Ontologies (*)																					
EthanPlants	spire.umbc.edu	owl (Full)	en	56838	56838	0	0.0 (0)																					
disease_ontology	loria.fr	owl (Lite)	en	19136	19136	0	0.0 (0)																					
chebi	loria.fr	owl (Lite)	en	13059	13059	0	0.0 (0)																					
NP	purl.org	owl (Lite)	en	6370	6370	1	0.0 (0)																					
circ	inkrmke.org	daml		2497	1740	757	0.0 (0)																					
PATO	purl.org	owl (Lite)	en	1895	1892	4	0.0 (0)																					
SO	purl.org	owl (Lite)	en	1525	1505	21	0.0 (0)																					
BIRNLex-Anatomy	purl.org	owl (DL or Full)		1398	1398	0	1.67 (2)																					
astronomy	archive.astro.umd.edu	owl (Full)		1231	1518	0	7.5 (7)																					
swpatho2	swpatho.ag-nbi.de	owl (DL or Full)	de	784	594	0	0.0 (2)																					
ccmd-Instrument	geobrain.laits.gmu.edu	owl (Lite)	de en	771	879	0	0.0 (0)																					
BIRNLex-OrganismalTaxonomy	purl.org	owl (DL or Full)		760	760	0	1.67 (2)																					
FonctionsCerveau	www-timc.imag.fr	owl (DL or Full)	en es fr	759	521	0	0.0 (0)																					
BuildingsAndPlaces	ordinancesurvey.co.uk	owl (Full)		684	677	0	0.0 (0)																					
cuosity	loria.fr	owl (Lite)		674	674	0	0.0 (0)																					
quality	loria.fr	owl (Lite)		653	653	0	0.0 (0)																					
facc	cwu.edu	owl (Lite)		603	603	0	0.0 (0)																					
PW	purl.org	owl (Lite)	en	601	600	2	0.0 (0)																					
neuron_ontology	purl.org	owl (Full)		528	1389	4	1.5 (1)																					
BIRNLex-Investigation	purl.org	owl (DL or Full)		516	516	0	0.0 (7)																					
instruments	archive.astro.umd.edu	owl (Full)		489	493	0	7.0 (5)																					
physics	archive.astro.umd.edu	owl (Full)		476	643	0	4.13 (5)																					
YAO	archive.astro.umd.edu	owl (DL or Full)		439	561	0	9.55 (4)																					
skeleton	harmonia.uni-klu.ac.at	owl (DL or Full)	de en	408	208	0	0.0 (0)																					
MGEDOntology	mgd.sourceforge.net	daml		338	228	110	0.0 (0)																					
BIRNLex-Disease	purl.org	owl (Full)		333	333	0	1.75 (3)																					
ccmd-platform	geobrain.laits.gmu.edu	owl (Lite)	en fr	329	417	0	0.0 (0)																					
unit	loria.fr	owl (Lite)		253	253	0	0.0 (0)																					
NMR	msi-ontology.sourceforge.net	owl (Full)		246	246	0	0.0 (2)																					
dfile_7_10_0E_11_30_27_AM	accessegov.org	owl (Full)	en fr pl	245	0	118	0.0 (0)																					

Figura 2.15. Librería de ontologías de OntoSelect.

2.1.6.1 Repositorios de ontologías biomédicas

Debido a la importancia de disponer de mecanismos que faciliten el acceso a la gran cantidad de ontologías biomédicas existentes, el dominio de la biomedicina es en el que se pueden encontrar mejores ejemplos de repositorios de ontologías.

Actualmente, el repositorio de ontologías biomédicas más popular es el conocido como Open Biomedical Ontologies (OBO), creado y mantenido por el Centro Nacional para la Ontología Biomédica (National Center for Biomedical Ontology, NCBO) de EE.UU., a través de su BioPortal (National Center for Biomedical Ontology, 2010). OBO contiene más de 200 ontologías que tratan diversos subdominios de la biomedicina.

Este volumen refleja el creciente uso de las “bio-ontologías” en tareas de anotación para la compartición y análisis de conjuntos de datos biológicos. Paralelamente al desarrollo de este repositorio, los desarrolladores de un subconjunto de las ontologías del repositorio OBO dieron lugar a la iniciativa conocida como OBO Foundry (Smith et al., 2007), un experimento colaborativo basado en la aceptación voluntaria de sus

participantes de un conjunto de principios²² que extienden los propuestos inicialmente por OBO, requiriendo además que las ontologías:

1. Sean desarrolladas de forma colaborativa.
2. Utilicen relaciones comunes definidas de forma no ambigua.
3. Proporcionen procedimientos para tener en cuenta retroalimentación (*feedback*) de los usuarios y para identificar versiones sucesivas.
4. Tengan un claro tema de estudio (de tal manera que, por ejemplo, una ontología dedicada a los componentes celulares no debería contener términos como *base de datos* o *train*).

En la figura 2.16 se muestra el aspecto del explorador del repositorio de ontologías biomédicas de BioPortal.

VERSION	AUTHOR	UPLOADED ON	GROUP	STATUS
1.0	Allen Institute for Brain Science	08/08/2009		Explore
1.0.23	Yongqun "Oliver" He	11/09/2010		Explore
1.101	Ghislain Atemezing	06/28/2009		Explore
1993	May Cheh	02/05/2010	UMLS	Explore
1.2 (inferred)	Nick Drummond, Georgina Moulton, Robert Stevens, Phil Lord	07/02/2010		Explore
1.8	David Blackburn	12/17/2010		Explore
See Remote Site	AmphiAnat list	11/02/2009		Explore
unknown	Http://animaldiversity Administrators	08/31/2010		Explore
0.1	Jeremy Espino	09/30/2010		Explore

Figura 2.16. Repositorio de ontologías OBO, accesible a través de BioPortal²³.

²² Los principios de OBO Foundry se pueden consultar en <http://obofoundry.org/>

²³ <http://bioportal.bioontology.org/ontologies/>

2.1.7 Evaluación y selección de ontologías

El importante incremento del número de ontologías disponibles a través de Internet y la aparición de repositorios de ontologías a gran escala ha provocado que cada vez sea más difícil para un usuario encontrar la ontología u ontologías más adecuadas para afrontar un problema determinado en el que interesa reutilizar ontologías ya existentes.

Esta situación ha motivado la aparición de métodos dirigidos a facilitar el proceso de selección de ontologías. Se trata de métodos semi-automáticos o automáticos que, a partir de un conjunto de ontologías almacenadas en un repositorio y de acuerdo a un conjunto de requerimientos o características del problema concreto que se desea afrontar (e.g. un conjunto de términos que se desean describir semánticamente), proporcionan como salida la ontología u ontologías más adecuadas para afrontar dicho problema.

Aunque la selección de ontologías es un área de investigación relativamente reciente (últimos 5-7 años), depende fundamentalmente de un proceso en el que diversos investigadores han realizado grandes esfuerzos en las últimas décadas: la evaluación de ontologías. La evaluación de ontologías consiste en medir la calidad de una ontología de acuerdo a un conjunto de criterios predefinidos. Se trata de un proceso complejo, cuyo origen se remonta a principios de la década de los 90, debido a la necesidad de disponer de estrategias que orientasen el proceso de desarrollo de ontologías. Más recientemente, en especial desde la aparición de la Web Semántica (Berners-Lee, et al., 2001), la evaluación de ontologías se ha orientado hacia la tarea de seleccionar la mejor ontología para una tarea determinada.

En dominios como la biomedicina, la cantidad y variedad (formatos, ubicaciones, etc.) de ontologías existentes es actualmente tan grande que elegir una ontología para una tarea de anotación o para el diseño de una aplicación específica constituye un reto importante. Además, la reusabilidad es una práctica deseable en el desarrollo de ontologías, debido a dos razones (Jonquet et al., 2010): (1) Construir una ontología desde cero es una tarea larga y compleja. (2) Es vital para la comunidad evitar la aparición de múltiples ontologías que compiten entre sí para representar el mismo conocimiento. Debido a esto, en dominios como el biomédico, la selección de ontologías se ha convertido en una prioridad en los últimos años.

Debido a que los procesos de evaluación y selección de ontologías son un aspecto crucial de esta tesis, éstos se explican de forma más detallada en el capítulo 3 (Estado de la Cuestión) de este documento, en el que también se analizan los principales trabajos al respecto.

2.2 Web 2.0 y conocimiento colectivo

Una de las ideas que se proponen en esta tesis es el uso de conocimiento elaborado de forma colectiva o (colaborativa) por múltiples usuarios, siguiendo el enfoque de lo que se conoce como Web 2.0. En este apartado, se explica en qué consisten las nociones de Web 2.0 y conocimiento colectivo.

2.2.1 De la Web 1.0 a la Web 2.0

La World Wide Web (WWW o simplemente Web) fue concebida por Tim Berners-Lee en 1989 a partir de un proyecto del CERN (Organización Europea para la Investigación Nuclear) y cambió radicalmente el modo en que la gente recopilaba información y accedía a ésta. Hoy en día, la WWW se ha convertido en un gigantesco repositorio de información en continuo e imparable crecimiento, que no ha dejado de ofrecer nuevas posibilidades y usos no previstos inicialmente, pudiendo considerarse sus repercusiones técnicas y sociales como una auténtica revolución a finales del siglo XX y comienzos del XXI.

El estallido de la burbuja tecnológica en el otoño de 2001 marcó un momento crucial para la Web. Mucha gente concluyó que la expectación sobre la Web era exagerada, cuando de hecho las burbujas y las consiguientes crisis económicas parecen ser una característica común de todas las revoluciones tecnológicas. Las crisis económicas marcan típicamente el punto en el cual una tecnología en ascenso está lista para ocupar su lugar en el escenario económico. Se descarta a los impostores, las historias de éxito verdaderas muestran su fortaleza, y comienza a comprenderse qué separa a los unos de los otros (O'Reilly, 2005).

El concepto de “Web 2.0” nació de una sesión de *brainstorming* realizada entre O'Reilly Media²⁴ y MediaLive International²⁵ cuya finalidad era proponer ideas para una nueva conferencia. Dale Dougherty²⁶, pionero de la Web y vicepresidente de O'Reilly Media, observó que lejos de “estrellarse”, la Web era más importante que nunca, con apasionantes nuevas aplicaciones y con sitios Web apareciendo con sorprendente regularidad. Y lo que es más, las compañías que habían sobrevivido al desastre parecían tener algunas cosas en común. ¿Podría ser que el derrumbamiento de las “punto-com” supusiera algún tipo de giro crucial para la Web, de tal forma que una llamada a la acción tal como la Web 2.0 pudiera tener sentido? Hubo acuerdo y así nació la conferencia de la Web 2.0²⁷.

Actualmente, una búsqueda en Google por el término “Web 2.0” devuelve más de 23 millones de resultados²⁸. Este dato es un claro reflejo de que el concepto se encuentra profundamente arraigado en Internet. Sin embargo, en algunos casos todavía existe desacuerdo en cuanto a lo que significa Web 2.0.

Como muchos conceptos importantes, la Web 2.0 no tiene unos límites claros, sino más bien, un núcleo gravitacional. Se puede visualizar la Web 2.0 como un sistema de principios y prácticas que conforman un verdadero sistema solar de sitios que muestran algunos o todos esos principios, a una distancia variable de ese núcleo.

La figura 2.17 muestra un “mapa meme²⁹“ de la Web 2.0 que fue desarrollado en una sesión de *brainstorming* durante el FOO Camp³⁰, una conferencia en O'Reilly Media. Este mapa manifiesta las principales ideas que irradian desde el núcleo de la Web 2.0.

²⁴ <http://oreilly.com/>

²⁵ <http://www.medialiveintl.com/>

²⁶ Dale Dougherty es uno de los cofundadores (junto a Tim O'Reilly) de O'Reilly Media

²⁷ <http://www.web2summit.com/>

²⁸ Búsqueda realizada por última vez el 8 de noviembre de 2010.

²⁹ “meme” es un neologismo inventado por Richard Dawkins (etólogo británico, zoólogo, teórico evolutivo y divulgador científico nacido en 1941) por su semejanza fonética al término gen (en inglés) y que se refiere a la información mínima acumulada en nuestra memoria y captada generalmente por imitación (mimesis), por enseñanza o por asimilación.

³⁰ <http://wiki.oreillynet.com/foocamp05/>

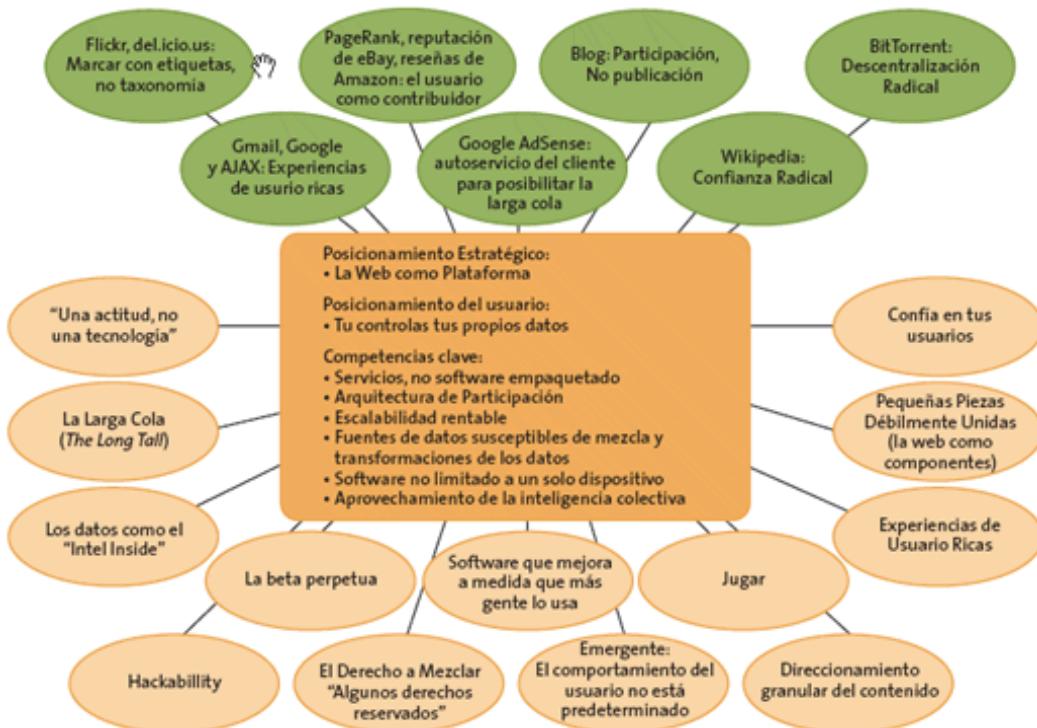
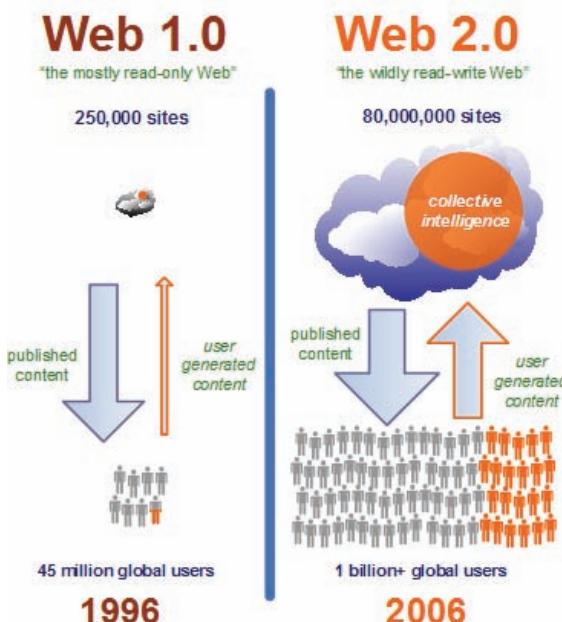


Figura 2.17. Mapa MEME de la Web 2.0 (O'Reilly, 2005).

Figura 2.18. Diferencia entre la generación y consumo de contenidos en Web 1.0 y Web 2.0. Extraída de Examiner.com³¹.

³¹ <http://www.examiner.com/web-2-0-in-kansas-city/web-1-0-versus-web-2-0/>

Se define como Web 1.0 a la antecesora de la 2.0 aunque esto no quiere decir que haya desaparecido, es más, muchas de sus tecnologías siguen siendo utilizadas hoy en día coexistiendo con las nuevas tecnologías sociales. La figura 2.18 muestra la diferencia entre la generación y el consumo de contenidos en la Web 1.0, considerada de “sólo lectura”, en la que existen pocos productores de contenidos y muchos consumidores de los mismos, y la Web 2.0 o de “lectura y escritura”, en la que los usuarios se transforman en productores que colaboran para generar contenidos. A continuación se describen las características más importantes de ambos enfoques:

Las principales características de la Web 1.0 son las siguientes:

- Pocos productores de contenidos.
- Muchos lectores de esos contenidos.
- Páginas estáticas.
- La actualización de los sitios Web no se realiza de forma periódica.
- Sitios direccionales y no colaborativos.
- Los usuarios son lectores consumidores.
- Interacción mínima reducida a formularios de contacto, inscripción, boletines, etc.
- Discurso lineal emisor-receptor.

Mientras que en la Web 2.0:

- Los usuarios se transforman en productores de contenido.
- Web colaborativa.
- Posibilidad de publicar las informaciones y realizar cambios en los datos sin necesidad de conocimientos tecnológicos avanzados.
- Facilita las interacciones.
- Facilita la publicación, la investigación y la consulta de contenidos Web.
- Información en permanente cambio.

- Software gratuito o de muy bajo coste.

A pesar de que hoy en día los sitios Web suelen combinar características de la Web 1.0 y de la Web 2.0, se puede percibir la diferencia entre ellas a través de ejemplos de cada caso (ver tabla 2.1).

Tabla 2.1. Ejemplos de Web 1.0 frente a sus equivalentes Web 2.0. Traducida de (O'Reilly, 2005).

Web 1.0	Web 2.0
DoubleClick	Google AdSense
Ofoto	Flickr
Akamai	BitTorrent
Mp3.com	Napster
Britannica Online	Wikipedia
Sitios Web personales	Blogging
Evite	Upcoming.org and EVDB
Especulación con nombres de dominios	Optimización de motores de búsqueda
Páginas vistas	Coste por click
Screen scraping	Servicios Web
Publicación	Participación
Content Management Systems	Wikis
Directorios (taxonomías)	Etiquetado (folksonomías)
Stickiness	Sindicación

2.2.2 Conocimiento colectivo

Como se ha visto, la Web 2.0 o Web social está representada por un tipo de sitios Web y aplicaciones en las cuales la participación del usuario es la clave de su valor. Es habitual utilizar el término “inteligencia colectiva”, “conocimiento colectivo” o “sabiduría de las masas” para hacer referencia al valor creado de forma colaborativa a través de múltiples contribuciones individuales de personas que escriben artículos para Wikipedia, publican fotos etiquetadas en Flickr o comparten sus enlaces en Del.icio.us. El potencial de compartición de conocimiento actual es incomparable históricamente (Gruber, 2008). Nunca antes una cantidad tan grande de personas con creatividad y conocimiento habían tenido la oportunidad de ponerse en contacto a través de una red universal. El resultado, a día de hoy, es un increíble volumen de información y diversidad de perspectivas, y una cultura de participación en masa que da soporte a una inagotable fuente de contenidos.

Como explica Thomas Gruber (Gruber, 2008) La clave del conocimiento colectivo es una adecuada sinergia entre humanos y máquinas. Las personas son productores y consumidores: son la fuente del conocimiento, y tienen problemas e intereses reales. Las máquinas son los facilitadores: almacenan, recuerdan, buscan y combinan datos, y llevan a cabo inferencias lógicas. Las personas aprenden a través de la comunicación con otras personas, y crean nuevo conocimiento en el contexto de tal conversación. Internet hace posible a las máquinas ayudar a crear nuevo conocimiento y aprender unos de otros de forma más efectiva.

Con la llegada de la Web 2.0, se ha alcanzado una situación en la que millones de usuarios ofrecen su conocimiento de forma *online*, haciendo posible el almacenamiento, búsqueda y compartición del conocimiento de forma sencilla. La tecnología ha hecho posible la creación de sistemas de conocimiento colectivo, facilitando las siguientes tareas (Gruber, 2008):

- **Captura.** El bajo precio de los sensores, microprocesadores, memoria, redes de fibra óptica y telefonía móvil ha facilitado el acceso a mucha gente a computadoras, teléfonos móviles, cámaras digitales y banda ancha. Esto ha permitido a la gente compartir su información de forma digital y pasar más tiempo *online*.
- **Almacenamiento.** Los bajos precios de los discos de almacenamiento, tanto a nivel casero como en grandes servidores, ha eliminado las barreras para compartir grandes cantidades de información.
- **Distribución.** Internet es un superconductor de información que mantiene a todo el planeta conectado.
- **Comunicación.** Los sistemas de colaboración asíncrona (e.g. e-mail, wikis, blogs, etc.) superan las barreras espaciales y temporales, así como el número de individuos que toman parte de una conversación. Actualmente se puede hablar con cualquiera y aprender de las conversaciones de otros sin estar físicamente cerca de ellos.

2.3 La Web Semántica

La Web ha cambiado drásticamente la disponibilidad de la información accesible de forma electrónica. Actualmente, 14 billones de páginas web³² se encuentran indexadas por los principales buscadores, y son consultadas por un gran número de usuarios en constante evolución. Como muestra la figura 2.19, en los últimos 15 años los usuarios de Internet han aumentado desde 16 millones en 1995 hasta los más de 1.650 millones actuales.

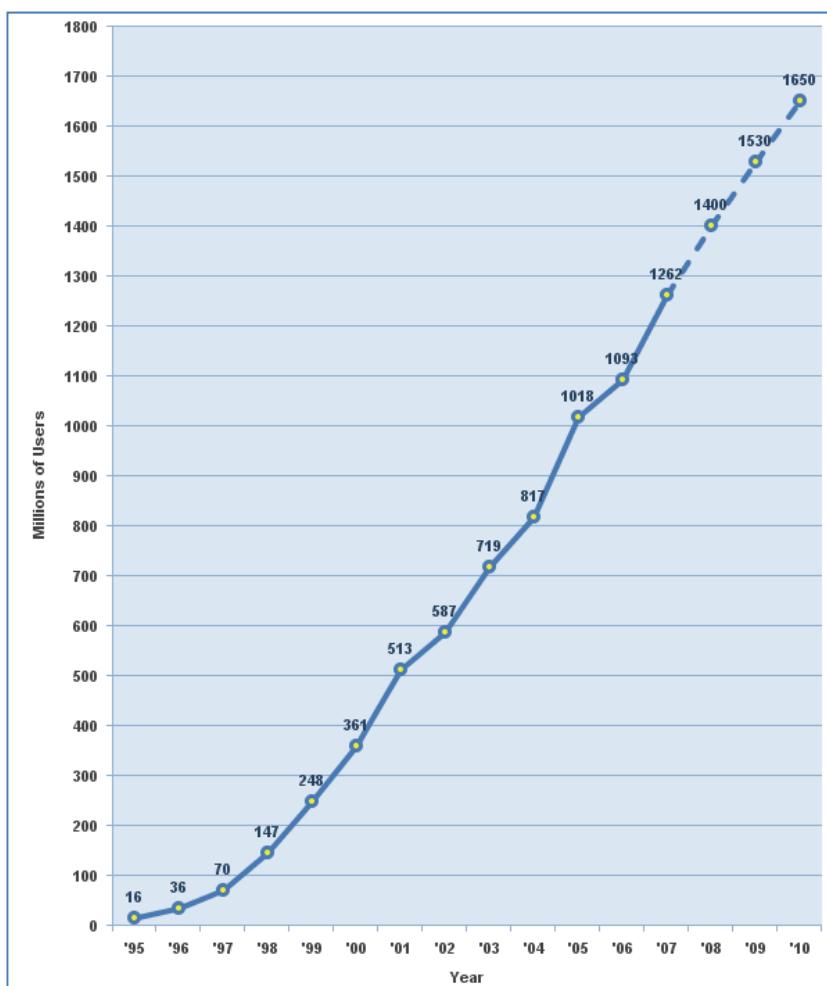


Figura 2.19. Evolución del número de usuarios de Internet desde 1995 hasta 2010. Extraída de Internet World Stats³³.

³² <http://www.worldwidewebsize.com/>

³³ <http://www.internetworldstats.com/>

Esta inmensa cantidad de datos ha provocado que cada vez sea más difícil realizar búsquedas, acceder, presentar y mantener la información relevante. Esto se debe a que el contenido de la información se presenta, ante todo, en lenguaje natural. La Web ha evolucionado como medio para el intercambio de información entre personas, no entre máquinas. Como consecuencia, el contenido semántico (i.e. el significado de la información en las páginas Web), se codifica de tal forma que únicamente es comprensible para los seres humanos. La figura 2.20 ejemplifica esta situación mediante un sitio Web para la predicción del tiempo. Se puede ver que la presentación de los datos en el buscador es fácilmente interpretable para los seres humanos. Sin embargo, para que una computadora pudiese procesar estos datos de forma automática, tendría que entender el significado de los conceptos de temperatura, humedad, visibilidad, etc.

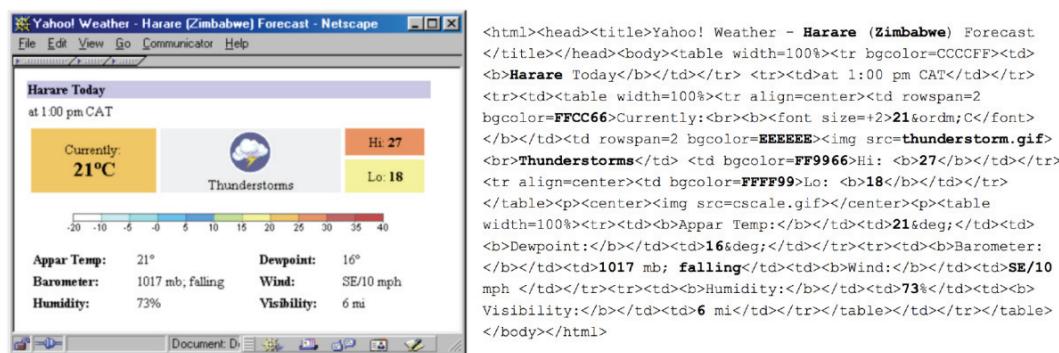


Figura 2.20. Sitio Web para la predicción del tiempo y código HTML correspondiente. Se puede observar que la Web está orientada a los seres humanos, no a las máquinas (Castells, 2003).

Para resolver los mencionados problemas de la Web actual, se pensó en describir los recursos Web de forma que las máquinas puedan comprenderlos. Se trata de dotar de semántica a los contenidos. Esta evolución de la Web es lo que se conoce como Web Semántica.

Realmente, la idea de la Web Semántica parecía estar en la cabeza de Berners-Lee desde el principio, como se desprende de la lectura de su obra “*Weaving the Web*” (Berners-Lee, 1999). Pero la Web fue evolucionando sin tener demasiado en cuenta el aspecto semántico de la información. Sin embargo, Tim Berners siguió trabajando en esta idea. En 1998 publica un borrador para el W3C titulado “*Semantic Web Road Map*” en dónde explica el concepto de Web Semántica (Berners-Lee, 1998). Para muchos,

este momento se considera el “nacimiento” de la Web Semántica. Sin embargo, será “*The Semantic Web*”, un artículo aparecido en *Scientific American* en Mayo de 2001 (Berners-Lee, et al., 2001) y del que fueron coautores Tim Berners-Lee, James Hendler y Ora Lassila, el que dará a conocer definitivamente la idea de Web Semántica. En este artículo se establecen diversos escenarios imaginarios en los que agentes software son capaces de realizar numerosas tareas accediendo al contenido de diferentes páginas de la WWW. Los autores señalan que, para que este escenario sea factible, debería cambiar la manera de representar contenido en la Web (hasta ahora diseñado para que los seres humanos puedan leerlo) para incluir una “semántica bien definida” que permitiese a componentes software acceder al mismo.

La Web Semántica pretende resolver los problemas de semántica implícita, el caótico crecimiento de los recursos y la ausencia clara de organización de la actual Web. Para esto, propone clasificar, proporcionar estructura, y anotar los recursos con semántica explícita (metadatos semánticos) procesable por las máquinas. La figura 2.21 ilustra esta propuesta. En la actualidad, la Web se puede ver como un grafo formado por nodos de un único tipo (páginas HTML), y aristas (hipervínculos) entre los que no existe diferencia. De esta manera, por ejemplo, para una máquina no existe distinción entre un sitio Web personal de un pintor y una tienda online de arte, y los enlaces que llevan a las páginas con información sobre las clases de un profesor no son diferentes de aquéllos que dirigen hacia sus publicaciones. Sin embargo, en la Web Semántica cada nodo (recurso) posee un tipo (o clase, o categoría) específico (e.g. profesor, tienda, pintor, libro, etc.), y las aristas representan relaciones explícitamente diferenciadas (pintor - pintura, profesor - departamento, libro - editorial, etc.).

En definitiva, la Web Semántica no es una nueva Web independiente de la actual, es una extensión de la Web existente, en la que la información se ofrece con un significado bien definido, permitiendo a computadoras y personas trabajar de forma cooperativa (Berners-Lee, et al., 2001). Se trata de una Web de información legible por las computadoras y de servicios automatizados que amplían considerablemente las capacidades de la actual Web.

La Web Semántica también se conoce como el núcleo de la Web 3.0, como evolución de la Web 2.0 o Web social, aunque los principales expertos en la materia no

son muy partidarios de usar versiones para referirse a la Web (Lassila & Hendler, 2007).

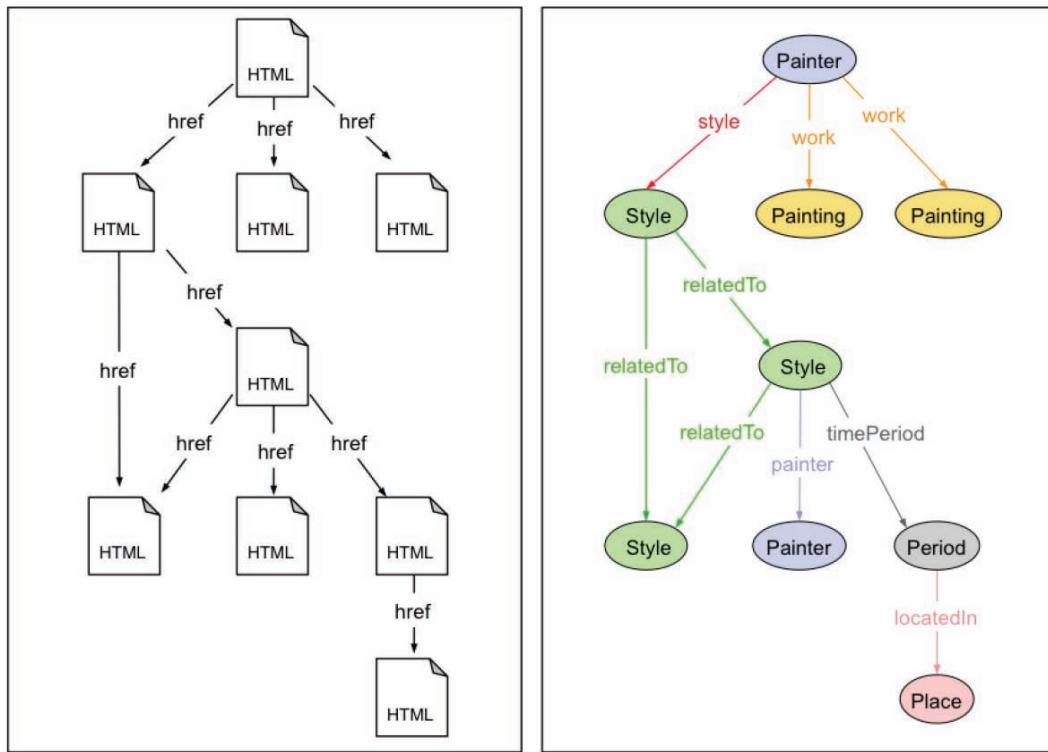


Figura 2.21. Estructuración de contenidos en la actual Web (izquierda) vs. el mismo contenido estructurado de acuerdo a la Web Semántica (derecha).

2.3.1 ¿Por qué es necesaria una Web Semántica?

La Web Semántica no está dirigida únicamente a la World Wide Web, sino que representa un conjunto de “tecnologías” que trabajarán igual de bien en “intranets” corporativas internas. La Web Semántica proporcionará solución a varios problemas clave a los que se enfrentan las actuales arquitecturas de las tecnologías de la información. A continuación, se comenta brevemente en qué consiste cada uno de ellos (Daconta et al., 2003; Davies et al., 2003):

La sobrecarga de información. Es un problema del que los expertos en tecnologías han advertido desde hace 50 años y que, hoy en día, necesita una solución. Por supuesto, está generalmente aceptado que el problema de la sobrecarga de la información, en forma de datos y noticias, ha empeorado considerablemente con la propagación de Internet, el correo electrónico y, actualmente, con la mensajería

instantánea. Desafortunadamente, la tendencia de la sociedad hacia la producción en lugar de hacia la reutilización del conocimiento ha dejado este problema sin resolver hasta que, finalmente, ha alcanzado unas dimensiones preocupantes. Se espera que la Web Semántica ayude a resolver el problema de la sobrecarga de información, facilitando y agilizando el tratamiento de grandes volúmenes de datos. A continuación, se comentan algunas debilidades de los sistemas de administración de conocimiento que la Web Semántica intentará solventar:

- *Búsqueda de la información.* Las búsquedas existentes, basadas en palabras clave, muchas veces recuperan información irrelevante que incluye ciertos términos de diferentes significados. Además, se pierde información cuando se utilizan diferentes términos con el mismo significado sobre el contenido deseado. Tradicionalmente, la recuperación de la información se centra en la relación entre una consulta dada (o perfil de usuario) y el almacén de información. Por otra parte, la explotación de las interrelaciones existentes entre diferentes piezas de información (facilitada mediante la utilización de ontologías) puede introducir en un contexto significativo información que previamente se encontraba aislada.
- *Extracción de la información.* Actualmente, es el ser humano el que está obligado a consultar y leer múltiples fuentes de información para extraer la información relevante de las mismas. Esto es así porque no existen agentes automáticos que posean el conocimiento requerido para recuperar tal información a partir de representaciones textuales, y fallan en la integración de información distribuida en diferentes fuentes.
- *Mantenimiento de la información.* El mantenimiento de fuentes de texto poco estructuradas es una actividad difícil y costosa en tiempo cuando dichas fuentes son de gran tamaño. Mantener estas fuentes en estado coherente, correcto y actualizado requiere representaciones mecanizadas de semánticas que ayuden en la detección de anomalías.
- *Generación automática de documentos.* Si las computadoras fuesen capaces de generar conceptos de forma automática, podrían existir sitios Web adaptativos, que se reconfigurarían dinámicamente de acuerdo a perfiles de usuario u otros

aspectos de relevancia. La generación de presentaciones de información semi-estructuradas a partir de datos semi-estructurados requiere una representación de la semántica de las fuentes de información que sea accesible por las computadoras.

Los sistemas aislados. Un sistema aislado es un sistema en el que todos los componentes se encuentran interconectados para su único trabajo conjunto. Por lo tanto, la información sólo fluye en el interior de estos sistemas y no puede ser compartida con otros sistemas u organizaciones que la necesiten. Por ejemplo, el cliente únicamente puede comunicarse con un software intermedio específico que únicamente conoce a una base de datos simple de esquema fijo. Este tipo de sistemas son producto de soluciones propietarias, adquiridas de forma gradual, e integradas mediante métodos *ad hoc*. Se trata de sistemas no extensibles, no interoperables y que poseen muchas funciones duplicadas. Es necesario migrar desde estos sistemas a la próxima generación de sistemas de información, un conjunto de sistemas extensibles, interoperables y mantenibles. Las tecnologías de la Web Semántica pueden ayudar a resolver este problema, acabando con los sistemas aislados. Concretamente, las tecnologías de la Web Semántica serán especialmente efectivas en la eliminación de los sistemas aislados de base de datos.

La baja agregación de contenidos. Integrar información procedente de diferentes fuentes es un problema recurrente en un gran número de áreas, como la agregación de cuentas financieras, la agregación de datos epidemiológicos, la agregación de portales, las compras comparativas, y la minería de datos. Desafortunadamente, la técnica más utilizada para llevar a cabo estas actividades es el *screen scraping*, que consiste en que un programa (denominado *screen scraper*) extrae datos (en formato texto) de la salida por pantalla de otro programa y los utiliza para realizar otra tarea (e.g. proporcionarlos como entrada a otro programa). Esta técnica se fundamenta en que es menos costoso (y menos arriesgado) construir un programa que emule el comportamiento de un usuario, que integrar los sistemas existentes para que interactúen adecuadamente. Cuando el *screen scraping* se realiza en la Web, se denomina *Web scraping*.

2.3.2 Arquitectura de la Web Semántica

La arquitectura de la Web Semántica se basa en una pila o pirámide de estándares propuesta por Tim Berners-Lee en el año 2000, durante una conferencia en el evento XML 2000 (ver figura 2.22).

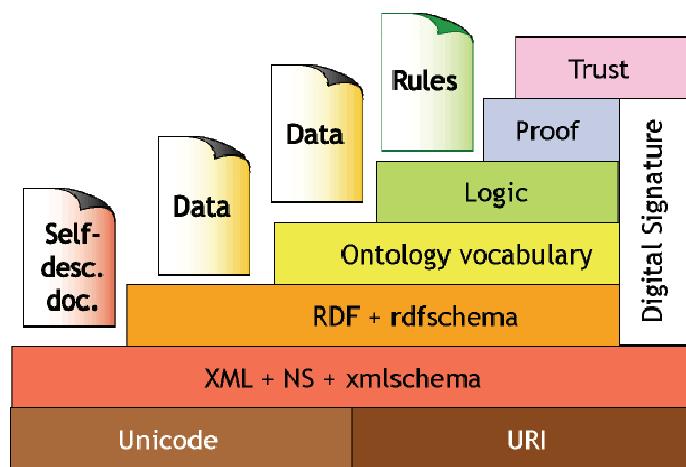


Figura 2.22. Pila de estándares de la Web Semántica. Presentada por Tim Berners-Lee durante una conferencia en el evento XML 2001 (ver <http://www.w3.org/2000/Talks/1206-xml2k-tbl/slides10-0.html>).

A continuación, y en base a los trabajos de Greenberg y colegas, y Davies y colegas (Davies, et al., 2003; Greenberg et al., 2003), se proporciona una breve explicación sobre las capas de dicha arquitectura.

- **URI.** URI es el acrónimo de *Uniform Resource Identifier* (identificador uniforme de recursos). Una URI es un texto corto que identifica únicamente cualquier recurso (servicio, página, documento, dirección de correo electrónico, etc.) accesible en una red. Es importante distinguir entre URI y URL (*Uniform Resource Locator* o localizador uniforme de recursos). El concepto de URL queda englobado por el de URI y está cada vez más en desuso. Normalmente un URI consta de dos partes, un identificador del método de acceso al recurso o protocolo, por ejemplo, `http:`, `mailto:` o `ftp:`, y el nombre del recurso, por ejemplo “`//www.udc.es`”. De esta forma, URI’s serían `mailto:marcosmartinez@udc.es` o `http://www.udc.es`, que también es un ejemplo de URL.

- **Unicode** es un estándar industrial cuyo objetivo es proporcionar el medio por el cual un texto en cualquier forma e idioma pueda ser codificado para el uso informático. Unicode proporciona un número único para cada carácter, sin importar la plataforma, el programa o el idioma. Unicode se ha vuelto el más extenso y completo esquema de codificación de caracteres, siendo el más dominante en la internacionalización y adaptación local del software informático.
- **XML + NS + XML Schema.** XML (Bray et al., 2000) son las siglas de *eXtensible Markup Language* (lenguaje de marcado ampliable o extensible). Desarrollado por el World Wide Web Consortium (W3C), es una versión simple de SGML (*Standard Generalized Markup Language* o lenguaje de marcado generalizado). El lenguaje XML es la base de un número de actividades de desarrollo de software cada vez mayor. Está diseñado para permitir el marcado de documentos de estructura arbitraria, al contrario que HTML (Raggett et al., 1999), el cual se diseñó para documentos de hipertexto de estructura fija. Un documento XML bien formado crea un árbol equilibrado de conjuntos anidados de etiquetas de apertura y de cierre, cada una de las cuales puede incluir varios pares atributo-valor. No hay un vocabulario fijo de etiquetas o conjunto de combinaciones permitidas, así que éstas se pueden definir para cada aplicación particular. El lenguaje XML, y más recientemente, los esquemas XML (*XML schemas*) facilitan la creación, utilización e interoperabilidad sintáctica de los vocabularios de metadatos. Los espacios de nombres (*namespaces*, NS), los cuales se identifican mediante URIs, aseguran la interoperabilidad semántica entre vocabularios de metadatos.
- **RDF y RDF Schema.** RDF (Lassila & Swick, 1999) o (*Resource Description Framework*, Infraestructura para la Descripción de Recursos) es una recomendación del W3C desarrollada para describir recursos Web mediante metadatos, sin embargo, RDF también es adecuado para representar datos. El modelo de datos RDF consiste en tres componentes: recursos u objetos, propiedades o atributos y declaraciones o valores. El bloque básico en RDF es una terna “recurso–propiedad–valor”. El modelo de datos RDF no dispone de mecanismos para definir las relaciones entre propiedades y recursos. Debido a

esto surge la Descripción de Vocabulario RDF (*RDF Vocabulary Description*), también conocido como *RDF Schema* (esquema RDF) o RDFS (Brickley & Guha, 2003). Existe otro término, RDF(S), que se refiere a la combinación de RDF y RDFS, que es ampliamente utilizado como formato de representación en muchas herramientas y proyectos. Existen cantidad de recursos para trabajar con RDF(S), como exploradores, editores, validadores, etc. No se debe confundir RDFS con RDF(S). La capa de vocabulario de ontología se refiere a este lenguaje.

- **Vocabulario de ontología (*ontology vocabulary*)**. Esta capa representa la principal arteria de metadatos de la Web Semántica. Permite crear y registrar esquemas de descripción y clasificación de diferentes niveles de complejidad, para que los agentes puedan interpretar datos de forma inteligente, realizar inferencias y llevar a cabo tareas.
- **Lógica (*logic*)**. La capa lógica permite la escritura de reglas de conocimiento. Por ejemplo: si N denota nuevo correo en una bandeja de entrada de correo electrónico, entonces si una N aparece para un mensaje particular, el mensaje se considera correo no leído. Esta inferencia se basa en la evidencia proporcionada por la letra N . La capa lógica de la Web Semántica trabaja sobre este principio básico a través de Lógica de Predicados de Primer Orden, aunque puede admitir otros tipos de lógica. Un agente puede alcanzar una conclusión durante la realización de una tarea basada en hechos presentados a partir de metadatos codificados semánticamente.
- **Prueba y confianza (*proof and trust*)**. Son las dos últimas capas horizontales, y están asentadas sobre la capa lógica. La capa de prueba realiza una validación de la evidencia procedente de la actividad lógica inferencial. Por otra parte, la capa de confianza está relacionada con la integridad de la prueba, cuyo origen se puede determinar descendiendo hacia las capas inferiores de la pila de estándares. La funcionalidad de estas dos capas es altamente dependiente de la creación de metadatos correctos y de confianza.
- **Firma digital (*digital signature*)**. La Web Semántica dispondrá de mecanismos que permitan a los usuarios confiar en la seguridad de las

operaciones que realizan y en la calidad de la información proporcionada. La firma digital podría ayudar a validar la integridad de los metadatos que los agentes utilizan para razonar y completar tareas.

2.3.3 El papel de las ontologías en la Web Semántica

Las ontologías son una tecnología clave en la futura Web Semántica, puesto que permiten entrelazar el conocimiento humano con su procesamiento por parte de las computadoras. El tipo de ontología más común para la Web, actualmente, dispone de una taxonomía y de un conjunto de reglas de inferencia.

La taxonomía define clases de objetos y relaciones entre ellos. Por ejemplo, una dirección se puede definir como un tipo de *emplazamiento*, y se pueden definir *códigos de ciudad*, aplicables únicamente a *emplazamientos*. Clases, subclases y relaciones entre entidades son una herramienta muy potente para la utilización en la Web. Se puede expresar un gran número de relaciones entre entidades por medio de la asignación de propiedades a clases y permitiendo a las subclases heredar tales propiedades. Si *códigos de ciudad* debe ser de tipo *ciudad* y las ciudades normalmente poseen página Web, se puede consultar el sitio Web asociado a un *código de ciudad* incluso sin la necesidad de disponer de una base de datos que enlace un código de una ciudad con un sitio Web.

Las reglas de inferencia de las ontologías proporcionan todavía más potencia. Una ontología puede expresar la regla “*Si un código de ciudad está asociado con un código de comunidad autónoma, y una dirección utiliza tal código de ciudad, entonces esa dirección posee el código de comunidad autónoma asociado*”. Un programa podría deducir, por ejemplo, que la Universidad de A Coruña, que está en A Coruña, se encuentra en Galicia, la cual está en España, y por tanto se deben utilizar los estándares de formateo de datos relativos al estado español.

Una vez la Web empiece a poblararse de páginas con contenido ontológico, comenzarán a surgir diversas soluciones a los actuales problemas de terminología (y otros) (Berners-Lee, et al., 2001). El significado de términos o códigos XML utilizados en una página Web se puede definir por medio de punteros desde la página Web a una ontología. Por supuesto, los mismos problemas que existían antes surgirán si una persona apunta a una ontología que define que las *direcciones* contienen un *código postal* y

otra apunta a una ontología que utiliza códigos de otro tipo en las direcciones. Este tipo de confusión se puede resolver estableciendo relaciones de equivalencia entre las entidades de las ontologías, es decir, mediante la utilización de lo que se conoce como “alineamiento de ontologías”.

Las ontologías pueden aumentar la funcionalidad de la Web actual de muchas formas. Se pueden utilizar simplemente para mejorar la precisión de las búsquedas Web, pues el programa de búsqueda podría buscar únicamente aquellas páginas que se refieren a un determinado concepto. Aplicaciones más avanzadas podrán utilizar las ontologías para relacionar la información de una página con las estructuras de conocimiento y reglas de inferencia asociadas a ella. Sin embargo, el gran poder de la Web Semántica se hará notable cuando se comiencen a desarrollar programas que recopilen contenidos Web a partir de varias fuentes, procesen esta información e intercambien los resultados con otros programas. La efectividad de estos agentes software se incrementará exponencialmente a medida que aumente la cantidad de información legible por las computadoras y los servicios automáticos en la Web.

2.3.4 Web Semántica vs. Web 2.0

La Web Semántica se encuentra inevitablemente ligada a la Web 2.0. El propio Tim Berners-Lee planteaba su visión de la Web Semántica diciendo que *“no es una Web diferente, sino una extensión de la Web actual, en la que la información se proporciona con un significado bien definido, permitiendo a las computadoras y personas trabajar de forma cooperativa”* (Berners-Lee, et al., 2001).

Tom Gruber explica que la Web Semántica juega un papel esencial para hacer posible la creación de sistemas de conocimiento colectivo, pues permite *“Crear un nuevo valor para los datos recopilados”*. Gruber explica que éste es el principal papel de la Web Semántica en lo que respecta a los sistemas de conocimiento colectivo (Gruber, 2008). Aunque a día de hoy existen múltiples formas de crear conocimiento colectivo mediante la agregación de contribuciones de usuarios individuales, existen pocos ejemplos que vayan más allá, resumiendo u ordenando los datos. Existen dos principales maneras en que la tecnología de Web Semántica puede ser útil al respecto:

1. Se puede añadir valor a los datos añadiendo datos estructurados. Es decir, las tecnologías de Web Semántica pueden añadir datos estructurados relacionados con el contenido de las contribuciones del usuario de forma que hagan posible un procesamiento más sencillo de los mismos.
2. Los estándares y la infraestructura de la Web Semántica pueden permitir la compartición y procesamiento de datos a través de aplicaciones de Web 2.0 independientes y heterogéneas. Mediante la combinación de datos estructurados y no estructurados, extraídos de diferentes sitios en Internet, la tecnología de Web Semántica puede proporcionar la base para el descubrimiento de nuevo conocimiento que no se encuentra contenido en ninguna fuente actual, y la solución a problemas no previstos por los creadores de sitios Web individuales.

Tanto la noción de Web 2.0 como el de Web Semántica, corresponden a estados evolutivos de la Web, y la Web Semántica correspondería en realidad a una evolución posterior, a la Web 3.0 o Web Inteligente (ver figura 2.23). La combinación de sistemas de redes sociales como Facebook³⁴ o Twitter³⁵, y el uso de conjuntos de etiquetas (o *tags*) en “blogs” y “wikis” (folksonomías), confieren a la Web 2.0 un aire semántico sin serlo realmente. Sin embargo, en el sentido más estricto, para hablar de Web Semántica, se requiere el uso de estándares de metadatos como Dublin Core³⁶, y en su forma más elaborada, de ontologías y no de folksonomías.

Por lo tanto, se puede ver la Web Semántica como una forma de Web 3.0. Existe una diferencia fundamental entre ambas versiones de Web (2.0 y Semántica), y es el tipo de participante y las herramientas que se utilizan. La 2.0 tiene como principal protagonista al usuario humano, que escribe artículos en su blog o colabora en un wiki. El requisito es que, además de publicar en HTML, emita parte de sus aportaciones en diversos formatos para compartir esta información como son los RSS³⁷, ATOM³⁸, etc. mediante la utilización de lenguajes estándares como XML³⁹. La Web Semántica, sin embargo, está orientada hacia el protagonismo de agentes software capaces de procesar

³⁴ <http://www.facebook.com/>

³⁵ <http://twitter.com/>

³⁶ <http://dublincore.org/>

³⁷ <http://feed2.w3.org/docs/rss2.html/>

³⁸ <http://www.atomenabled.org/>

³⁹ <http://www.w3.org/XML/>

información representada utilizando ontologías (e.g. en OWL⁴⁰) y llevar a cabo tareas de razonamiento.

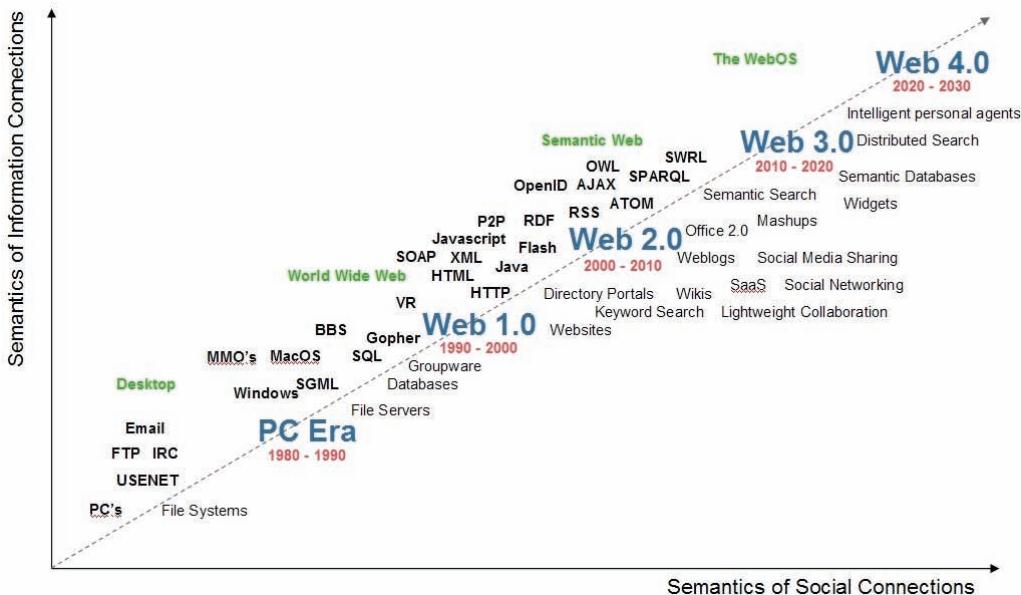


Figura 2.23. Principales conceptos y variedades tecnológicas en la evolución del Web, a lo largo del tiempo. Los ejes representan el grado de semántica de las conexiones de información y de las conexiones sociales.

2.4 Metadatos y anotación semántica

2.4.1 Metadatos y metadatos semánticos

El primer paso para comprender qué es la anotación semántica de los datos es entender qué son los metadatos. Según el Committee on Cataloging: Description and Access (CC:DA), de la American Library Association (ALA), los metadatos se definen de la forma siguiente (CC:DA, 1999):

“Los metadatos son datos estructurados y codificados, que describen características de entidades de información para ayudar a identificar, descubrir, evaluar y administrar dichas entidades”.

Otra definición es la que proporciona Kenneth Haase en 2004 (Haase, 2004) en la que define metadato como:

⁴⁰ <http://www.w3.org/TR/owl-features/>

“Cualquier dato que expresa conocimiento sobre un elemento sin requerir examinar el elemento en sí mismo”.

En otras palabras, los metadatos son “datos sobre los datos”, esto es, información sobre la información misma, datos descriptivos de otros datos.

Así, algunos ejemplos de metadatos sobre diferentes tipos de datos serían:

- Para un **libro**: el título, autor, fecha de publicación, tema, identificador (e.g. ISBN: *International Standard Book Number*), tamaño, número de páginas, idioma del texto, etc.
- Para una **fotografía**: fecha y hora a la que fue tomada, detalles y ajustes de la cámara utilizada (e.g. enfoque, apertura, exposición), etc.
- Para un archivo de **audio**: nombre del artista y álbum, título de la canción, género, año, compositor, número de pista, etc.

Como explica Corcho (Corcho, 2006), los metadatos se pueden asociar a documentos, que pueden encontrarse disponibles en formato electrónico (e.g. PDF, DOC, HTML) o en formatos tradicionales (e.g. papel), públicamente (e.g. en la Web) o bien almacenados de forma privada (e.g. disco duro de un PC). Los metadatos también se pueden asociar a aplicaciones, que pueden estar ejecutándose en un PC o en la Web de forma pública (e.g. en forma de Servicios Web). Y también pueden encontrarse metadatos que describen el contenido de recursos de almacenamiento de datos como las bases de datos (BDs).

En cuanto a la representación de los metadatos, éstos se pueden expresar en gran variedad de lenguajes y utilizando diversos vocabularios, y pueden representarse en diferentes formatos, ya sea electrónica o físicamente, y pueden ser creados y mantenidos utilizando diferentes tipos de herramientas (desde editores de texto a herramientas de generación de metadatos), automáticas o manuales.

Mientras el valor económico de los metadatos aumenta a medida que crece el volumen de contenidos disponibles (ya que éstos facilitan el acceso y recuperación de dichos contenidos), la media del valor económico de estos contenidos debe, de la misma manera, disminuir a medida que aumenta dicho volumen de contenidos (a más contenidos, menos valiosos individualmente). Estas dos tendencias se representan en la figura 2.24. El potencial de esta transformación es enorme y motiva la inversión en

nuevas variedades tecnológicas y métodos de creación, gestión y mantenimiento de los metadatos.

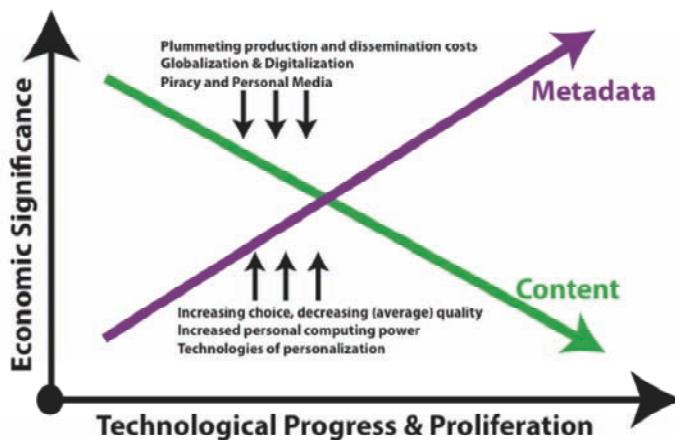


Figura 2.24. Importancia económica de los metadatos vs. contenido (Haase, 2004).

En definitiva, los metadatos pueden describir cualquier tipo de información, sea cual sea el soporte en el que se encuentre almacenada, y existen diversas opciones de creación y mantenimiento de estos metadatos. Sin embargo, el potencial de explotación de la información descrita a través de los metadatos adquiere especial relevancia cuando se trata de grandes volúmenes de información, que están accesibles de forma pública y los metadatos se encuentran almacenados electrónicamente en formatos estándar, de tal manera que pueden ser procesados de forma (semi)automática por las computadoras.

2.4.2 Calidad de los datos y metadatos semánticos

Cuando se trabaja con metadatos, es muy importante tener en cuenta la calidad de éstos. Los metadatos de calidad permiten discernir perfectamente entre contenidos relevantes e irrelevantes, permitiendo ahorrar tiempo y esfuerzo a los usuarios.

Una de las principales medidas de calidad de los metadatos es la precisión (*precision*): los metadatos deben ser lo suficientemente precisos como para permitir distinguir de forma efectiva entre diferentes elementos, sin la necesidad de que una persona los examine. Además, la precisión es importante para determinar qué elementos deben incluirse o excluirse, ya sea por motivos de irrelevancia, redundancia, o porque no sean

apropiados. En cualquier caso, cuanto más precisos sean, más valiosos serán los metadatos.

Existen dos importantes problemas con la precisión de los metadatos (Haase, 2004). El primero de ellos se refiere a que, en bases de datos y modelos de metadatos convencionales, aumentar la precisión de las descripciones reduce la efectividad (*recall*) global. Por ejemplo, si una imagen se ha descrito utilizando el metadato *bicicleta*, una persona que esté buscando imágenes de *medios de transporte* no será capaz de encontrarla. El segundo problema hace referencia a que para conseguir producir metadatos de calidad, una persona debe invertir un esfuerzo equiparable al que se trataba de evitar mediante el uso de metadatos.

El problema de la reducción de efectividad (*bicicleta* vs. *medio de transporte*) se puede resolver en gran medida utilizando un tipo especial de metadatos, que permitirían expresar la relación existente entre *Bicicleta* y *Medio de transporte*. Este tipo de metadatos se conoce como *metadatos semánticos*, se definen como “*metadatos que enlazan los términos relacionados unos con otros*” (Haase, 2004), y “*han sido concebidos para ser leídos, comprendidos y procesados por máquinas*” (Jiang et al., 2008). Además, “*los metadatos semánticos describen contenidos en base a información específica del dominio*” (Sheth et al., 2002). Por ejemplo, en un dominio financiero, metadatos semánticos podrían ser la empresa, nombre de la empresa, ejecutivos, etc. Si se trata del dominio deportivo, podrían ser jugador, equipo, liga, etc. Cuando se habla de metadatos semánticos, se está haciendo referencia a metadatos que han sido definidos utilizando conceptualizaciones consensuadas, o vocabularios compartidos. En la actualidad, las estructuras que proporcionan el contexto para los metadatos semánticos son las ontologías.

2.4.3 ¿Qué es la anotación?

El diccionario de la Real Academia Española proporciona las siguientes definiciones:

“*Anotación: Acción y efecto de anotar*”

“*Anotar: Poner notas en un escrito, una cuenta o un libro*”

Sin embargo, no existe una única definición de anotación comúnmente aceptada, sino que existen diferentes propuestas, de múltiples autores. A continuación, se revisan algunas de estas definiciones:

En 1998, Marshall definió dos diferentes dimensiones en la anotación (Marshall, 1998): la *anotación informal* y la *anotación formal*. La *anotación informal* incluiría anotaciones como las notas personales escritas al margen de un libro mientras se lee un artículo, y es el tipo de anotación que concuerda con las definiciones previamente presentadas. Por otra parte, la *anotación formal* tendría que ver con metadatos que siguen estándares estructurales y valores asignados en base a nomenclaturas consensuadas. En este caso, el uso de modelos conceptuales consensuados, representados utilizando lenguajes estándar de representación del conocimiento, como las ontologías, se ubicaría al final del espectro de formalidad.

Otra de las clasificaciones de anotación más conocidas es la Bechhofer y colegas (Bechhofer et al., 2002) quienes, cuatro años más tarde, restringen la anotación a tres tipos: textual, de enlace y semántica. La *anotación textual* sería el proceso de añadir comentarios o notas a un texto. Se trata del tipo de anotación que se ha estado utilizando durante muchos años en dominios como la Biología, especialmente en bases de datos biológicas, siendo útiles, por ejemplo, para describir secuencias de proteínas usando anotaciones. La *anotación de enlace* extiende la anotación textual añadiendo enlaces a las anotaciones sobre el contenido, aparte del texto. Finalmente, la *anotación semántica* es aquélla en la que el contenido de la anotación contiene información semántica, extraída de ontologías, y es el tipo de anotación en la que se centrará esta tesis.

Más recientemente, en 2007, Shah y Musen (Shah & Musen, 2007) aclaran los dos diferentes tipos de anotación más comunes en el dominio biomédico. Estos autores señalan que en este dominio es muy importante tener clara la diferencia entre las anotaciones de aserción (*assertion annotations*) y las anotaciones de metadatos (*metadata annotations*). La diferencia entre estos dos tipos de anotación se explica con un ejemplo en la figura 2.25.

Las anotaciones de aserción son aquellas aserciones o declaraciones acerca de relaciones entre entidades biológicas y los procesos en los cuales participan. Los biólogos utilizan este tipo de anotaciones para expresar verdades en el dominio. Un ejemplo de anotación de aserción sería la afirmación: “El gen DMP53 está asociado con la muerte celular”. Las anotaciones de metadatos, en cambio, proporcionan información adicional (i.e. metadatos) acerca de un experimento o un conjunto de

datos relacionados con una entidad biológica. Un ejemplo de anotación de metadatos sería un texto descriptivo de una imagen de un microarray, o un concepto de una ontología que representa lo existente en dicha imagen. Las primeras se utilizan para la representación explícita de conocimiento, mientras que las segundas son útiles para la indexación de contenidos. En este tipo de anotaciones, las anotaciones de metadatos, serán a las que se hará referencia en el resto de este documento. Más concretamente, se hablará de anotaciones de metadatos semánticos, las cuales se tratarán con más detalle en el siguiente apartado.

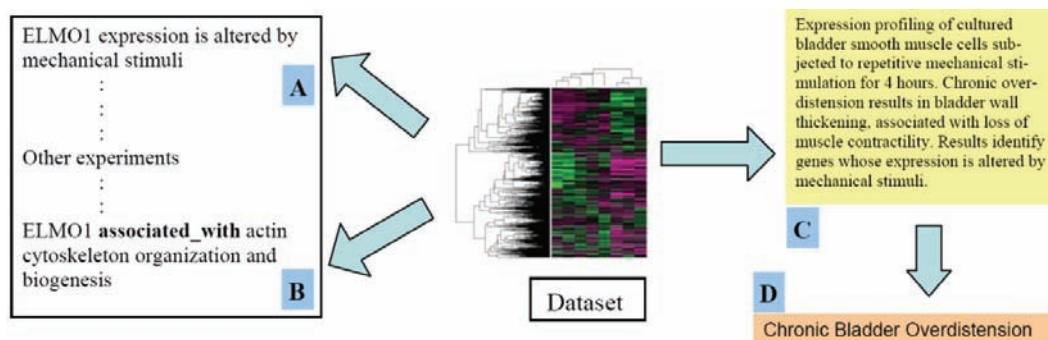


Figura 2.25. Diferencia entre anotaciones de aserción y anotaciones de metadatos. Los datos son analizados por un investigador, que extrae una afirmación determinada a partir de ellos (A). Otro investigador, en base a múltiples afirmaciones de este tipo, construye una anotación de aserción (B). Por otra parte, las descripciones realizadas por el investigador sobre los datos son las anotaciones de metadatos, las cuales se pueden realizar en lenguaje natural (C), o bien en base a un vocabulario controlado u ontología (D) (Shah & Musen, 2007).

2.4.4 Anotación semántica

Como explica Kiryakov (Kiryakov et al., 2004), no existe una única definición aceptada para la anotación semántica, sino que existen diversas definiciones de diferentes autores sobre este término. Incluso a veces, diferentes autores utilizan distintos términos para referirse al proceso de anotación semántica (e.g., *etiquetado semántico* (Dill et al., 2003), *anotación de metadatos* (Scerri et al., 2005)). A continuación, se discutirán algunas de las principales definiciones de anotación semántica.

Una de las definiciones más generales de anotación semántica es la proporcionada por Handschuh en 2003, en la que define la anotación semántica como “*el acto de decorar datos ya existentes con metadatos semánticos, los cuales describen los datos*” (Handschoh & Staab, 2003).

Sin embargo, existen otras definiciones más restrictivas, bien en cuanto al tipo de contenidos a anotar (e.g. textos, páginas Web, etc.), como a la formalidad de los metadatos utilizados. A continuación, se revisan algunas de estas definiciones:

Una de ellas, proporcionada por Kiryakov y colegas en 2004, restringe la anotación a contenidos textuales y al uso de ontologías como referencia para las anotaciones. Estos autores definieron la anotación semántica como “*el proceso que consiste en asignar a las entidades de un texto, enlaces a sus descripciones semánticas*” (ver figura 2.26) (Kiryakov, et al., 2004), considerando que estas descripciones semánticas son elementos de ontologías.

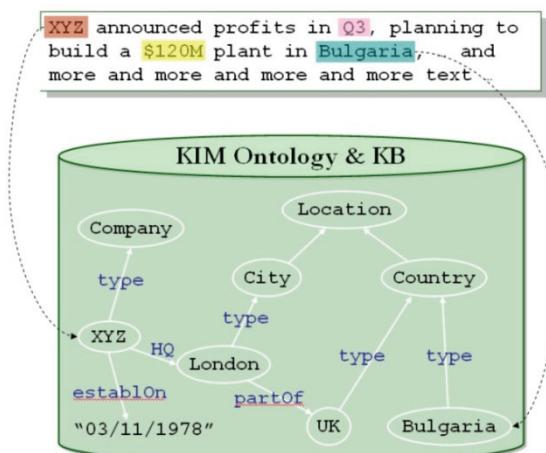


Figura 2.26. Ejemplo de anotación semántica de un texto. En este caso, los metadatos utilizados en la anotación son elementos de una ontología (Kiryakov, et al., 2004).

También existe un conjunto de definiciones que hacen referencia a anotación formal (en base a ontologías) pero restringen la anotación a contenidos Web, en el marco de la Web Semántica. Así, por ejemplo, Ding (2005) definió la anotación semántica como “*un proceso para etiquetar contenidos de páginas Web explícitamente, formalmente, y sin ambigüedad, utilizando ontologías*” (Ding, 2005). En ese mismo año, Scerri y colegas definen este término como “*el proceso de asociar descripciones semánticas a recursos Web, vinculándolos a clases y propiedades de ontologías*” (Scerri, et al., 2005). Para ellos, existen dos categorías de anotación: la *anotación interna*, en la que las anotaciones se incluyen dentro de los propios documentos HTML, y la *anotación externa*, cuando los metadatos se almacenan en una localización diferente, pero se encuentran vinculados a los contenidos HTML mediante enlaces. Otra definición de anotación semántica centrada

en la Web es la de Kudelka y colegas, quienes en 2006 definieron la anotación semántica como “*el proceso de añadir semántica formal (metadatos, conocimiento) al contenido Web, con el propósito de conseguir un acceso y una gestión más eficiente de estos contenidos*” (Kudelka et al., 2006).

Finalmente, Uren y colegas no restringen la anotación semántica a contenidos Web, sino que se refieren a documentos en general, pero también defienden la necesidad de expresar la semántica de manera formal. Ellos explican en 2006 (Uren et al., 2006) que “*la anotación semántica va más allá de anotaciones textuales informales sobre el contenido de un documento*”, como por ejemplo “*en esta sección se resumen las principales características estructurales de una célula...*”, “*los resultados podrían ir aquí*”, etc. Indican que estos tipos de anotación informal son comunes en aplicaciones como los procesadores de texto, y están dirigidas principalmente a su comprensión por parte del propio autor del documento. Sin embargo, afirman que “*cuando se habla de anotación semántica se está haciendo referencia a la identificación de conceptos y de relaciones entre conceptos en recursos de información, y que este tipo de anotación está principalmente concebida para su utilización por parte de las máquinas*”.

En cuanto al aspecto de la formalidad de la anotación semántica, tras revisar estas anotaciones se puede ver que algunos autores se refieren explícitamente al uso de ontologías (Ding, 2005; Kiryakov, et al., 2004; Scerri, et al., 2005), mientras que otros no son tan restrictivos, y hacen referencia a estructuras semánticas formales en general (Handschoh & Staab, 2003; Kudelka, et al., 2006; Uren, et al., 2006). Por otra parte, otros autores prefieren utilizar el término *anotación semántica basada en ontologías* para referirse a la anotación en la que los metadatos son elementos de ontologías (Erdmann et al., 2000; Jonquet et al., 2008; Khelif et al., 2007; Yesilada et al., 2003).

En la actualidad, la anotación semántica se está utilizando en diversos dominios y aplicaciones. Algunos ejemplos son la gestión del conocimiento médico y biomédico (French et al., 2009; Möller et al., 2009), gestión de galerías de imágenes en sitios Web como Flickr⁴¹ (Xu et al., 2009) o de vídeo (Min et al., 2009), o incluso la descripción de datos en dominios como la agricultura (Macario & Medeiros, 2009) o la arqueología (Karmacharya et al., 2009).

⁴¹ <http://www.flickr.com/>

Teniendo en cuenta que, a día de hoy, las ontologías se pueden considerar las estructuras de representación del conocimiento compartido y formal por excelencia, en esta tesis se considerará *anotación semántica* como sinónimo de *anotación semántica basada en ontologías*, entendiendo por ésta la anotación en la que únicamente se usan ontologías como modelos de referencia.

2.4.4.1 Anotación semántica basada en ontologías

Una ontología describe elementos del mundo real de forma comprensible y procesable por las máquinas. Además, los elementos que conforman una ontología y sus relaciones han sido acordados por una comunidad de personas con interés y conocimiento en el dominio descrito por la ontología.

Las estructuras ontológicas añaden un valor adicional a las anotaciones semánticas. Éstas abren un nuevo abanico de posibilidades sobre las anotaciones semánticas resultantes, como las posibilidades de inferencia o la navegación a través de conceptos, además de la referencia a un conjunto de conceptos comúnmente aceptados. Una ontología centra la atención del anotador en un conjunto predefinido de estructuras semánticas y, por lo tanto, constituye una guía sobre qué elementos de los recursos a anotar deben ser anotados, y cómo deben anotarse (Staab et al., 2001).

Sin embargo, aparte de las ventajas de la anotación semántica basada en ontologías respecto a la anotación basada en generación de metadatos semánticos menos formales (e.g. texto libre), este conjunto de capacidades también desencadena nuevos problemas que es necesario resolver. En particular, las interconexiones semánticas entre diferentes recursos deben ser gestionadas adecuadamente. Esto significa que una herramienta de anotación basada en ontologías debe tener en cuenta la *identidad de cada objeto* y gestionarla adecuadamente en varios recursos. Además, las ontologías pueden contener definiciones elaboradas de conceptos. Cuando su significado varía, o cuando determinados conceptos antiguos deben ser eliminados, o cuando surgen nuevos conceptos, la ontología cambia. Actualizar anotaciones previas es, por lo general, demasiado costoso, por lo que normalmente hay que abordar el problema del cambio de las ontologías en relación con sus anotaciones correspondientes. Finalmente, también es importante prevenir la anotación redundante, debida a la existencia de

recursos duplicados (e.g. páginas duplicadas en la Web) o al trabajo realizado por otros anotadores.

2.4.5 Beneficios de la anotación semántica

Como explican Uren y colegas (Uren, et al., 2006), la anotación semántica mejora las capacidades de recuperación de la información y de interoperabilidad de los sistemas en los que se utiliza.

En el caso de la anotación semántica basada en ontologías, la recuperación de la información se mejora debido a la habilidad de realizar búsquedas, que explotan la ontología para llevar a cabo inferencias sobre los datos procedentes de recursos heterogéneos, y permitiendo que componentes software autónomos (i.e., agentes) exploten estas anotaciones para recuperar información de forma inteligente y automática (Welty & Ide, 1999). Además, la anotación semántica basada en ontologías también permite resolver anomalías en las búsquedas, permitiendo llevar a cabo búsquedas basadas en conceptos, y evitando así confusiones debidas, por ejemplo, a palabras con varias acepciones (e.g. la palabra “estación” puede referirse al sitio donde hace parada un tren o un autobús, o a una de las partes en las que se divide el año, etc.).

Asimismo, la interoperabilidad se ve beneficiada, pues las anotaciones basadas en una ontología compartida proporcionan un punto de acceso común para la integración de información de fuentes heterogéneas. Para conseguir una buena interoperabilidad, una semántica bien definida es un requisito indispensable para asegurar que tanto el anotador como el consumidor de la anotación puedan compartir conocimiento. Así, por ejemplo, una anotación semántica podría relacionar la palabra “Roma” de un texto con varios elementos de una ontología, asociándola al concepto “Ciudad”, y a la instancia “Francia” de un concepto “País”, eliminando así cualquier ambigüedad acerca del significado de la palabra “Roma” mediante el conocimiento explícito y consensuado de la ontología.

Sin embargo, la anotación explícita de datos mediante conceptos de ontologías no es todavía una práctica común, debido principalmente a las siguientes razones (Jonquet, et al., 2008):

- La cantidad de ontologías relevantes está aumentando y conseguir acceder a todas ellas puede convertirse en una tarea pesada debido a sus diferentes formatos, localizaciones o APIs de acceso.
- Los usuarios no siempre conocen la estructura del contenido de una ontología o cómo utilizarla para llevar a cabo la anotación que necesitan.
- Anotar datos usando ontologías es una tarea muchas veces aburrida y adicional que no proporciona recompensa inmediata para el usuario.

3 Estado de la cuestión

La necesidad de disponer de estrategias de evaluación en el campo de las ontologías surgió en el año 1994, con los trabajos de Thomas Gruber y Asunción Gómez-Pérez en el Knowledge Systems Laboratory de la Universidad de Stanford (Gómez-Pérez, 1994; Gruber & Olsen, 1994).

Desde la concepción de la Web Semántica en 2001 (Berners-Lee, et al., 2001), el interés en la evaluación de ontologías se ha incrementado de forma considerable, dando lugar a diversos trabajos que proponen evaluar la calidad de una ontología de acuerdo a diferentes criterios como el número de clases, propiedades, la cantidad de niveles de la ontología, etc. (Supekar et al., 2004). Más recientemente, otros autores abordan diferentes formas de evaluar cómo de bien una ontología cubre un contexto o dominio particular (Alani et al., 2006; Alani, et al., 2007; Jonquet et al., 2009; Netzer et al., 2009; Vilches-Blázquez et al., 2009).

Este trabajo de investigación ha dado lugar a diversas aproximaciones para la evaluación de ontologías, que se pueden clasificar en dos grandes grupos atendiendo a la finalidad de la evaluación:

- **Aproximaciones dirigidas a mejorar el proceso de desarrollo de ontologías.** Este tipo de aproximaciones surgen a partir del trabajo de Gruber y Gómez-Pérez, y vienen motivadas por la necesidad de disponer de estrategias de evaluación de ontologías que permitan dirigir y corregir el proceso de construcción de las mismas.
- **Aproximaciones orientadas a medir la adecuación de una ontología para una tarea determinada.** Este conjunto de aproximaciones, más recientes y menos numerosas, surgen motivadas por la necesidad de disponer métodos que

permitan decidir cuál es la ontología u ontologías más adecuadas para una tarea determinada. En general, en este tipo de aproximaciones se asume que la ontología que se está evaluando ha sido construida correctamente. No se trata de evaluar la calidad de la ontología de forma independiente, sino su valía para afrontar una tarea determinada. En general, esta tarea es la descripción de recursos.

Los trabajos sobre evaluación de ontologías que se ubican en estos dos grandes grupos han dado lugar a varios términos vinculados a la tarea de “juzgar” ontologías y relacionados entre sí. Estos términos son: recuperación, búsqueda, evaluación, verificación, validación, valoración (*assessment*), ordenación, recomendación y selección de ontologías. Aunque una revisión literaria permite identificar este conjunto de términos, no existe un consenso en cuanto a sus definiciones. A veces, el mismo término se utiliza con significados diferentes, creando confusión.

En el siguiente apartado, se revisan las principales definiciones proporcionadas hasta el momento de estos términos. Posteriormente, se revisarán los principales trabajos relacionados en los ámbitos de la evaluación y selección de ontologías y se analizarán otras soluciones propuestas hasta la fecha para abordar el problema que se plantea.

3.1 Terminología y definiciones

Existen varios términos habitualmente utilizados en el ámbito del proceso de elegir la ontología u ontologías más adecuadas para un determinado contexto o problema. En la figura 3.1 se muestra un esquema de los principales términos en el área de la evaluación y selección de ontologías, basándose en los tipos de evaluación propuestos por Gómez-Pérez (Gómez-Pérez, 1994).

Los términos **búsqueda** y **recuperación** de ontologías están estrechamente relacionados, y hacen referencia al proceso de encontrar un conjunto de ontologías de un repositorio que satisfacen unos determinados requerimientos básicos (e.g. obtener todas las ontologías que contienen el concepto *cardiovascular disease*). Puesto que no todas las ontologías obtenidas proporcionarán el mismo nivel de satisfacción, medir

cómo de bien una ontología satisface la tarea para la que ha sido obtenida es parte de la evaluación de ontologías, que se explica a continuación.

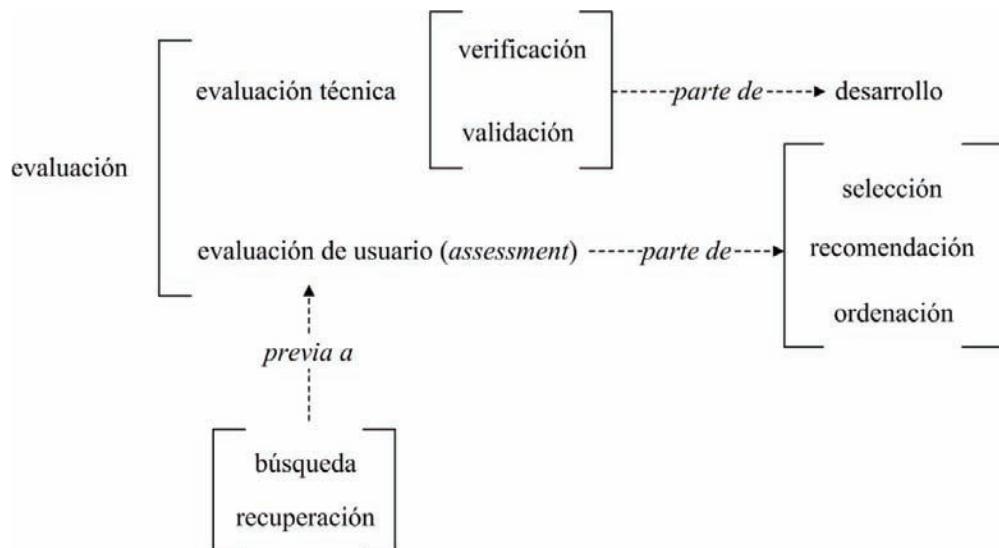


Figura 3.1. Esquema de los términos más utilizados en el ámbito de la evaluación y selección de ontologías.

3.1.1 Evaluación técnica y evaluación de usuario

En general, cuando se habla de **evaluación de ontologías**, es necesario distinguir entre evaluación técnica y evaluación de usuario.

La **evaluación técnica** consiste en juzgar técnicamente las características de la ontología respecto a un marco de referencia durante cada fase y entre las fases de su ciclo de vida (Gómez-Pérez, 1994). Ejemplos de marcos de referencia serían el mundo real, un conjunto de requisitos, o un conjunto de cuestiones de competencia (Gruninger & Fox, 1994). Se refiere a la correcta construcción del contenido de la ontología, es decir, a asegurar que las definiciones de la ontología (i.e. su estructura, contenido, sintaxis y propiedades semánticas), bien escritas en lenguaje natural, bien en lenguaje formal, implementan correctamente los requisitos de la ontología y las cuestiones de competencia, y garantizan que la ontología es coherente, completa, consistente y concisa (Gómez-Pérez, 1999). La evaluación técnica también incluye la evaluación del entorno software de desarrollo de la ontología y de su documentación (Gómez-Pérez, 1995). El objetivo es probar la conformidad del modelo del mundo real con el modelo modelado formalmente. En 2004, Gómez-Pérez (Gómez-Pérez, 2004)

sintetiza sus anteriores definiciones y define evaluación de ontologías (refiriéndose a la evaluación técnica) como “*un juicio técnico del contenido de la ontología con respecto a un marco de referencia (que puede ser: especificaciones de requisitos, cuestiones de competencia o el mundo real)* durante cada fase y entre las fases de su ciclo de vida”. Explica que la evaluación técnica debe realizarse contemplando los siguientes elementos:

- Definiciones de individuos y axiomas.
- Colecciones de definiciones y axiomas declarados de forma explícita en la ontología.
- Definiciones importadas de otras ontologías.
- Definiciones que se pueden inferir a partir de otras definiciones y axiomas.

La evaluación técnica de una ontología abarca su verificación y validación, y se refiere a la actividad realizada por el equipo de desarrollo de ontologías (Gómez-Pérez, 2004). A continuación, se explica en qué consisten la verificación y la validación de ontologías.

Verificación de ontologías (*ontology verification*) se refiere a “*construir la ontología correctamente, es decir, asegurando que sus definiciones implementan correctamente los requerimientos de la ontología y sus cuestiones de competencia, o funcionan correctamente en el mundo real*” (Gómez-Pérez, 2004). La verificación debe realizarse a lo largo de todo el ciclo de vida de la ontología, y tiene tres objetivos (Gómez-Pérez, 1994):

- Determinar la corrección de definiciones y axiomas, analizando lo que la ontología define de forma explícita, no define o define incorrectamente.
- Determinar el ámbito de las definiciones y axiomas, revisando qué se puede inferir, no se puede inferir, o puede ser inferido incorrectamente.
- Demostrar un conjunto de atributos bien definidos en definiciones y axiomas.

Validación de ontologías (*ontology validation*) se refiere a si se ha construido la ontología correcta, o lo que es lo mismo, comprobar que la ontología que se ha construido sirve para resolver las tareas para las que fue diseñada. Consiste en comprobar “*si el significado de las definiciones de la ontología realmente modelan el mundo real para el cual la ontología se creó. El objetivo es probar que el modelo del mundo (si existe y es conocido) corresponde con el mundo modelado formalmente*” (Gómez-Pérez, 2004). Es decir, se refiere a la tarea de evaluación realizada por el usuario final y tiene que ver con comprobar si el

significado de las definiciones de la ontología realmente representan el dominio para el que fue creada. Se trata de un proceso iterativo que garantiza que la ontología desarrollada corresponde con lo que se espera de ella (Gómez-Pérez, 1994).

Por otra parte, existe otro tipo de evaluación, conocida como **evaluación de usuario o valoración (assessment)**, que tiene que ver con el uso de la ontología para las tareas del mundo real para las que fue diseñada (Gómez-Pérez, 1995). Este tipo de evaluación se centra en “*juzgar la comprensión, usabilidad, utilidad, calidad y portabilidad de las definiciones desde el punto de vista del usuario*” (Gómez-Pérez, 2004). Diferentes tipos de usuarios y diferentes tipos de aplicaciones requieren diferentes formas de valorar la ontología, lo cual hace necesario disponer de una caracterización de las ontologías desde el punto de vista del usuario, considerando los tipos de aplicaciones que usan ontologías, para determinar cuándo una ontología existente es apropiada para una determinada aplicación o no (Gómez-Pérez, 2004). En un primer momento, varios autores han estudiado diferentes criterios de alto nivel para la valoración de ontologías de forma manual. Un claro ejemplo es el trabajo de Guarino y Welty (Guarino & Welty, 2002). Sin embargo, la necesidad de disponer de métodos para determinar la adecuación de una tarea a una ontología de forma rápida y sin la intervención de expertos ha llevado al desarrollo de aproximaciones automáticas que cubren diferentes perspectivas y niveles de evaluación. Es en este tipo de evaluación, la valoración automática de ontologías, en el que se centra el desarrollo de esta tesis.

Es frecuente encontrar literatura en la que se hace referencia al proceso de evaluación técnica, o al proceso de valoración, usando el nombre de evaluación. Así, por ejemplo, Brank y colegas definen evaluación de ontologías como (Brank, et al., 2005) “*el problema de valorar una ontología dada desde el punto de vista de un criterio de aplicación particular, típicamente para determinar cuál de varias ontologías encaja mejor con un propósito determinado*”. Esta definición es cercana a la definición de valoración de Gómez-Pérez. En el caso del trabajo de Alani et al. (Alani, et al., 2007) se define el problema de la búsqueda en un repositorio de ontologías como la tarea de “*obtener una colección de ontologías del repositorio relevantes según la temática de la consulta*”. Indican que una vez encontradas las ontologías relevantes, la evaluación de ontologías “*consiste en medir cómo de bien cada ontología satisface un conjunto de criterios predefinidos*”.

3.1.2 Ordenación, recomendación y selección

No existe un consenso acerca del término que se debe utilizar para hacer referencia a la tarea de elegir la ontología u ontologías más adecuadas según una consulta, problema o contexto concreto. Sin embargo, se puede determinar que los términos más utilizados para referirse a esta tarea son: ordenación, recomendación y selección de ontologías. Aunque en muchos casos se tratan como equivalentes, algunos autores proporcionan diferentes matices sobre su significado, que se explican a continuación.

La **ordenación de ontologías** (*ontology ranking*) es un problema importante, complementario a la búsqueda de ontologías: después de que un motor de búsqueda encuentra las ontologías relevantes, necesita ordenarlas para indicar cuáles son más relevantes que otras (Alani, et al., 2007). La idea es ahorrar una gran cantidad de tiempo y esfuerzo, eliminando la necesidad de examinar en detalle cada ontología de forma manual para determinar cómo de bien satisface las necesidades del agente o ingeniero de conocimiento. Alani y colegas definen la ordenación de ontologías como la tarea de “*ordenar las listas de ontologías derivadas de acuerdo a su relevancia respecto a la consulta*” (Alani, et al., 2006). Algunos autores consideran que la ordenación de ontologías forma parte de un proceso más general, conocido como selección de ontologías.

Sabou y colegas (Sabou, et al., 2006b) definen la **selección de ontologías** como “*el proceso que permite identificar una o más ontologías o módulos de ontologías que satisfacen ciertos criterios*”, explicando que el proceso de selección de ontologías es, en esencia, una tarea de evaluación de ontologías. De hecho, la evaluación de ontologías consiste en evaluar una ontología, de forma independiente a otras, siguiendo un criterio específico. Cuando es necesario seleccionar la ontología que proporciona la mejor cobertura para un *corpus*⁴² dado, un prerequisito de la selección consiste en evaluar todas las ontologías consideradas en base a dicho criterio. Por lo tanto, la evaluación de ontologías es el núcleo de la selección de ontologías. En su trabajo, consideran la ordenación como una subtarea de la selección de ontologías.

⁴² La R.A.E. define *corpus* como “Conjunto lo más extenso y ordenado posible de datos o textos científicos, literarios, etc., que pueden servir de base a una investigación”.

Finalmente, existe otro término frecuentemente utilizado: **recomendación de ontologías**. Cantador y colegas (Cantador et al., 2007) utilizan este término con un matiz colaborativo, refiriéndose al proceso de selección de ontologías que explota opiniones de diferentes usuarios para elegir la ontología más adecuada. Sin embargo, existen otros autores que utilizan el término recomendación como sinónimo de selección. Un ejemplo es el trabajo de Jonquet y colegas (Jonquet, et al., 2010). En esta tesis, se considerará que la ordenación es un paso previo a la selección de ontologías, y se tratarán los términos selección y recomendación como sinónimos.

3.2 Caracterización de las ontologías y criterios de evaluación

El paso previo a pensar en analizar o evaluar cómo de buena es una ontología, es conocer en detalle todas sus posibles características. Eduard Hovy indicaba hace unos años (Hovy, 2002) que, a medida que el número de ontologías disponibles crece día a día, la necesidad de caracterizar las ontologías se vuelve cada vez más importante, afirmando que “*sólo cuando esto se haya hecho, la Ingeniería de Ontologías evolucionará desde un arte a una ciencia*”. Asimismo, Asunción Gómez-Pérez explicaba en 2004 (Gómez-Pérez, 2004) que, para poder determinar cuándo una ontología ya existente es apropiada para una determinada aplicación o no, es necesaria una caracterización de las ontologías desde el punto de vista del usuario.

Varios han sido los autores que han trabajado en caracterizar las ontologías. A continuación, se realiza un repaso de los principales trabajos realizados al respecto.

En 1996, Asunción Gómez-Pérez identifica y define formalmente varios criterios a seguir para la evaluación de una ontología (Gómez-Pérez, 1996). Estos criterios son los siguientes (Gómez-Pérez, 2004):

- **Consistencia** (*consistency*). De forma general, se puede decir que una ontología es consistente si no es posible obtener conclusiones contradictorias a partir de definiciones de entrada válidas.

- **Compleción** (*completeness*). La no compleción es un problema fundamental en ontologías. Una ontología se dice completa si (1) todo lo que se supone que debe estar en la ontología está en ella o puede inferirse a partir de ella y, además, (2) todas las definiciones de la ontología son completas.
- **Concisión** (*conciseness*). Una ontología es concisa si (1) no almacena definiciones innecesarias o inútiles, (2) no existen redundancias explícitas entre definiciones o términos, y si (3) no se pueden inferir redundancias a partir de otras definiciones o axiomas.
- **Extensión** (*expandability*). Se refiere al esfuerzo requerido para añadir nuevas definiciones a una ontología y más conocimiento a sus definiciones, sin alterar el conjunto de propiedades bien definidas ya existentes.
- **Sensibilidad** (*sensitivity*). Tiene que ver con cómo pequeños cambios en una definición pueden alterar el conjunto de propiedades bien definidas existentes.

Una explicación detallada de estos criterios, acompañada de ejemplos intuitivos, se puede encontrar en el trabajo de Gómez-Pérez (Gómez-Pérez, 2004).

Noy y Hafner (Noy & Hafner, 1997) identifican el conjunto de características principales de una ontología (ver figura 3.2), que agrupan en 8 dimensiones diferentes (características generales, del proceso de diseño, de la taxonomía, relacionadas con la estructura interna de los conceptos y sus relaciones, relacionadas con los axiomas, relacionadas con los mecanismos de inferencia, relacionadas con las aplicaciones construidas usando la ontología y contribuciones importantes).

Sin embargo, como explican Arpírez y colegas (Arpírez, et al., 2000), esta propuesta tiene el problema de que limita la caracterización de las ontologías a este conjunto de características, además de la falta de criterios para clasificar las ontologías de acuerdo a algunas de estas características. Otro enfoque diferente es el propuesto por el propio Arpírez y colegas (Arpírez, et al., 2000), que proponen una caracterización de las ontologías desde el punto de vista del usuario, para tratar de responder a cuestiones como: ¿en qué lenguajes se encuentra disponible la ontología?, ¿cuál es el coste de la infraestructura software y hardware que se requiere para usar la ontología?, ¿la ontología ha sido evaluada técnicamente?, etc.

General	The purpose the ontology was created for General or domain specific Domain (if domain specific) Easy integration possible into a more general ontology Size: Number of concepts, rules, links, and so on Formalism used Implementation platform and language, if done Publication, if done
Design process	How was the ontology built? Was there a formal evaluation?
Taxonomy	What is the general taxonomy organization? Are there several taxonomies, or is everything in the same one? What is in the ontology: things, processes, relations, properties? What is the treatment of time? What is the top-level division? How tangled or dense is the taxonomy?
Internal concept structure and relations between concepts	Do concepts have internal structure? Are there properties and roles? Are there other kinds of relation between concepts? How are part-whole relations represented?
Axioms	Are there explicit axioms? How are the axioms expressed?
Inference mechanism	How is reasoning done (if any)? What are some instances of going beyond first-order logic?
Applications	Retrieval mechanism User interface Application in which the ontology was used
Contributions	Major strengths and contributions Weaknesses

Figura 3.2. Caracterización de las ontologías propuesta por (Noy & Hafner, 1997).

Esta caracterización de ontologías se muestra en la figura 3.3. Como se puede observar, las características se agrupan en tres categorías: características identificativas, descriptivas y funcionales. Al respecto, los autores hacen énfasis en que: (1) Algunas de estas características pueden no ser válidas para caracterizar ciertas ontologías. (2) Es posible que el desarrollador de ontologías no conozca los valores de todas estas características. (3) Se trata de una lista de características inicial, que puede ser completada y mejorada con nuevas características. (4) Se han definido pensando en permitir la búsqueda de ontologías de forma fácil e intuitiva.

IDENTIFYING	<table border="0"> <tr> <td>ONTOLOGY</td><td>name, server-sites, mirror-sites, Web-pages, FAQs available, mailing lists. NL- descriptions, built date</td></tr> <tr> <td>DEVELOPERS</td><td>name, Web-page, e-mail, contact name, telephone, FAX, postal address.</td></tr> <tr> <td>DISTRIBUTORS</td><td>name, Web-page, e-mail, contact name, telephone, FAX, postal address.</td></tr> </table>	ONTOLOGY	name, server-sites, mirror-sites, Web-pages, FAQs available, mailing lists. NL- descriptions, built date	DEVELOPERS	name, Web-page, e-mail, contact name, telephone, FAX, postal address.	DISTRIBUTORS	name, Web-page, e-mail, contact name, telephone, FAX, postal address.						
ONTOLOGY	name, server-sites, mirror-sites, Web-pages, FAQs available, mailing lists. NL- descriptions, built date												
DEVELOPERS	name, Web-page, e-mail, contact name, telephone, FAX, postal address.												
DISTRIBUTORS	name, Web-page, e-mail, contact name, telephone, FAX, postal address.												
DESCRIPTIVE	<table border="0"> <tr> <td>GENERAL</td><td>type of ontology, subject, purpose, ontological commitments, list of higher level concepts, implementation status, on-line and hard-copy documentation</td></tr> <tr> <td>SCOPE</td><td>number of concepts representing classes, number of concepts representing instances, number of explicit axioms, number of relations, number of functions. number of class concepts at first, second and third levels, number of class leaves, average branching factor, average depth, highest depth level</td></tr> <tr> <td>DESIGN</td><td>building methodologies, steps followed, level of formality of the methodology, building approach, level of specification formality, knowledge sources, reliability of knowledge sources, knowledge acquisition techniques, formalism paradigms, integrated ontologies, languages in which the ontology is available</td></tr> <tr> <td>REQUIREMENTS</td><td>hardware and software support</td></tr> <tr> <td>COST</td><td>price of use, maintenance cost, estimated price of required software, estimated price of required hardware</td></tr> <tr> <td>USAGE</td><td>number of applications, list of main applications.</td></tr> </table>	GENERAL	type of ontology, subject, purpose, ontological commitments, list of higher level concepts, implementation status, on-line and hard-copy documentation	SCOPE	number of concepts representing classes, number of concepts representing instances, number of explicit axioms, number of relations, number of functions. number of class concepts at first, second and third levels, number of class leaves, average branching factor, average depth, highest depth level	DESIGN	building methodologies, steps followed, level of formality of the methodology, building approach, level of specification formality, knowledge sources, reliability of knowledge sources, knowledge acquisition techniques, formalism paradigms, integrated ontologies, languages in which the ontology is available	REQUIREMENTS	hardware and software support	COST	price of use, maintenance cost, estimated price of required software, estimated price of required hardware	USAGE	number of applications, list of main applications.
GENERAL	type of ontology, subject, purpose, ontological commitments, list of higher level concepts, implementation status, on-line and hard-copy documentation												
SCOPE	number of concepts representing classes, number of concepts representing instances, number of explicit axioms, number of relations, number of functions. number of class concepts at first, second and third levels, number of class leaves, average branching factor, average depth, highest depth level												
DESIGN	building methodologies, steps followed, level of formality of the methodology, building approach, level of specification formality, knowledge sources, reliability of knowledge sources, knowledge acquisition techniques, formalism paradigms, integrated ontologies, languages in which the ontology is available												
REQUIREMENTS	hardware and software support												
COST	price of use, maintenance cost, estimated price of required software, estimated price of required hardware												
USAGE	number of applications, list of main applications.												
FUNCTIONAL	description of use tools, documentation quality, training courses, on-line help, operating instructions, availability of modular use, possibility of adding new knowledge, possibility of delaying with contexts, availability of PSMs												

Figura 3.3. Caracterización de ontologías desde el punto de vista del usuario, propuesta por (Arpírez, et al., 2000). Las características consideradas indispensables se muestran en negrita.

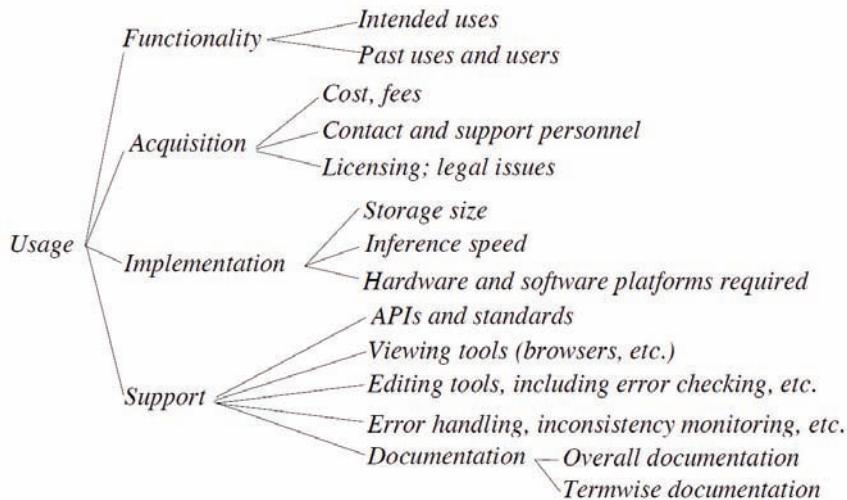


Figura 3.4. Taxonomía de características relacionadas con el “uso” de las ontologías, propuesta por (Hovy, 2002).

Hovy (Hovy, 2002) propone un gran número de características de las ontologías, clasificadas de acuerdo a tres taxonomías independientes. Estas taxonomías son las siguientes:

- 1) Forma. Se refiere a aspectos relacionados con la representación de la ontología.
Abarca aspectos teóricos y computacionales.
- 2) Contenido. Tiene que ver con los elementos (términos, instancias, reglas, etc.) utilizados para construir la ontología
- 3) Uso. Esta taxonomía clasifica aspectos de la ontología de interés para sus usuarios (e.g. coste, tamaño, documentación, etc.).

A modo de ejemplo, la taxonomía de “Uso”, se muestra en la figura 3.4.

3.2.1 Criterios de evaluación de ontologías biomédicas

A medida que el número de ontologías de diferentes dominios ha ido aumentando, también ha surgido la necesidad de disponer de criterios de evaluación adaptados a las características de las ontologías de dominios concretos, y a las necesidades específicas de reutilización de conocimiento en estos dominios. Un ejemplo es el campo de la biomedicina.

Recientemente, Tan y Lambrix (Tan & Lambrix, 2009) describen una lista de criterios a considerar cuando se desea seleccionar una ontología biomédica para tareas de minería de texto (*text mining*). Estos criterios son: el tipo de ontología (e.g. ontología de alto nivel, ontología del dominio, etc.), tecnología base (e.g. lenguaje de representación de la ontología, herramientas disponibles para trabajar con ella, etc.), variedad del conocimiento representado (e.g. conceptos, relaciones, instancias, etc.), cobertura del dominio y evaluación en un sistema real o contra un banco de pruebas (*benchmark*).

Sin embargo, el conjunto de criterios a tener en cuenta para la selección de ontologías biomédicas más exhaustivo hasta la fecha es el proporcionado por Jonquet y colegas desde el NCBO (Jonquet, et al., 2010). Se trata de un conjunto de 10 criterios funcionales, que se presentan a continuación:

1. **Automatización.** ¿La herramienta o método es completamente automático o requiere de interacción con el usuario que necesita la ontología?
2. **Dinamismo.** ¿La herramienta es lo suficientemente rápida para ser invocada de forma dinámica por aplicaciones cliente, y lo suficientemente

precisa para evitar requerir de intervención humana que realice una limpieza final de los resultados? Este criterio únicamente es aplicable a aproximaciones completamente automáticas.

3. **Correspondencias de términos.** ¿La herramienta se basa en algún tipo de búsqueda de correspondencias entre los términos de consulta? Este tipo de *matching* puede ser exacto o difuso, y en general se utiliza para evaluar la cobertura del contexto para una ontología.
4. **Correspondencias de propiedades.** ¿Se explota algún tipo de correspondencias entre los términos de consulta y los valores de propiedades de las clases (e.g. definiciones, sinónimos, etc.)?
5. **Expansión de consultas.** ¿Se realiza algún tipo de expansión de consultas para recuperar un conjunto de términos más representativo que permita mejorar la búsqueda de correspondencias con la ontología (e.g. usando recursos externos como WordNet o Wikipedia)?
6. **Medidas estructurales.** ¿La aproximación se basa en alguna medida formal de la estructura de la ontología (i.e. que utilice de alguna manera las relaciones de un concepto con otros)?
7. **Conectividad.** ¿Se explotan de alguna forma las referencias (e.g. importación, instancia, etc.) o vínculos (e.g. *mappings*) entre ontologías para dar mayor importancia a ontologías de referencia?
8. **Desambiguación.** ¿Se realiza algún tipo de desambiguación, ya sea para identificar a qué conceptos hace referencia la consulta de entrada, o bien para buscar las correspondencias entre los términos y la ontología?
9. **Razonamiento.** ¿La herramienta utiliza algún tipo de razonamiento?
10. **Popularidad.** ¿Se utiliza algún tipo de información por parte de los usuarios, ya sea directa (e.g. revisiones, notas) o indirecta (e.g. *logs* de uso, referencias en recursos Web) para ordenar las ontologías?

3.3 Requerimientos de la selección de ontologías

Sabou y colegas (Sabou, et al., 2006b) proporcionan un conjunto de requerimientos para la selección de ontologías en base a las características de los repositorios de ontologías más recientes. Estos requerimientos se resumen a continuación.

El gran tamaño de las librerías de ontologías requiere métodos de selección **automáticos y capaces de tratar con varias ontologías simultáneamente**. Esto también implica que los métodos de selección deberían tener un **buen rendimiento**. Este requisito también viene impuesto porque las herramientas basadas en ontologías, idealmente, utilizarán selección en tiempo de ejecución.

El método de selección también debe ser capaz de **tratar con colecciones heterogéneas** de ontologías. También hay que tener en cuenta el aspecto de la modularización. Las actuales aplicaciones semánticas requieren que los mecanismos de selección sean capaces de **proporcionar un módulo de una ontología en lugar de una ontología completa**.

Además, es necesario que el método de selección sea capaz de **proporcionar una combinación de ontologías como resultado**. Finalmente, también resulta importante que los métodos de selección sean capaces de **sacar provecho a la información proporcionada por las relaciones** de la ontología, no sólo tener en cuenta sus conceptos.

3.3.1 Selección manual vs. automática

Sabou y colegas (Sabou, et al., 2006a) explican las diferencias entre los requerimientos del proceso de selección de ontologías en un contexto en el que el ser humano puede actuar como mediador y en un contexto en el que es necesaria una selección automática.

En el caso de tareas en las que puede existir un ser humano como mediador:

- No es un problema si las ontologías devueltas no cubre el 100% de los términos de entrada (cobertura parcial), pues el usuario puede extender la ontología de acuerdo a sus necesidades.

- Se pueden admitir errores de correspondencias entre los términos de entrada y los conceptos de la ontología (cobertura imprecisa), pues el usuario puede realizar un filtrado de los errores.
- La respuesta del sistema no tiene porqué ser inmediata. El usuario puede esperar unos minutos para disponer de ontologías reutilizables, pues es un tiempo despreciable comparado con el necesario para desarrollar una ontología desde cero.

Sin embargo, en entornos en lo que se necesita disponer de reutilización de conocimiento automática, los requerimientos son mucho más estrictos:

- Es necesario que el resultado proporcionado por el sistema de selección (una o varias ontologías) proporcionen una cobertura completa de la tarea dada.
- No se admiten correspondencias erróneas entre los términos de entrada y los conceptos de la ontología.
- Es necesario proporcionar resultados en tiempo de ejecución, por lo que una respuesta rápida es vital.

Aunque ambicioso, el objetivo de disponer de métodos de selección que sean capaces de trabajar en un contexto de reutilización automática del conocimiento, es en el que se deben centrar los esfuerzos, ya que será el entorno en el que estos métodos serán necesarios la Web Semántica.

3.4 Aproximaciones existentes

Desde el nacimiento de las áreas de evaluación y selección de ontologías, han ido surgiendo diferentes aproximaciones para abordar estos problemas, basadas en diferentes criterios y con diferentes motivaciones. A lo largo de los años, varios han sido los autores que han tratado de clasificar las aproximaciones existentes en cada momento, tratando de proporcionar una visión general de este ámbito de investigación.

Una clasificación de referencia es la de Gómez-Pérez (Gómez-Pérez, 1994), ya explicada en la sección 3.1, que distingue entre dos tipos de evaluación: la evaluación técnica y la evaluación de usuario o valoración (*assessment*).

Otra clasificación, posterior, es la de Brewster y colegas (Brewster, et al., 2004), que clasifica los tipos de evaluación en tres categorías:

1. Evaluación de una ontología desde la perspectiva de los principios utilizados en su construcción. Indican que aunque estos principios de diseño son válidos a nivel teórico, es extremadamente difícil construir test automáticos que evalúen de forma comparativa varias ontologías en cuanto a su uso consistente de “criterios de identidad” o rigor de su taxonomía. Esto se debe a que tales principios dependen de conocimiento externo, que únicamente poseen los seres humanos.
2. Evaluar la efectividad de una ontología particular en el contexto de una aplicación. Consiste en evaluar la ontología en el marco de una serie de aplicaciones establecida.
3. Evaluar cómo de bien se ajusta una ontología a un dominio de conocimiento determinado. Puesto que no es posible evaluar directamente si una ontología encaja con el conocimiento que una persona posee sobre un dominio, una opción para llevar a cabo este tipo de evaluación es construir una representación estándar de dicho dominio o *gold standard* y comparar la ontología con él.

Otro de los trabajos más populares, más específico que los anteriores, es el de Brank y colegas (Brank, et al., 2005). Estos autores proporcionan dos clasificaciones diferentes para las aproximaciones de evaluación de ontologías, basadas en dos criterios distintos. Por una parte, según el método de evaluación seguido, se distinguen cuatro tipos de aproximaciones:

1. Basadas en comparar la ontología con un estándar (*gold standard*), que puede ser una ontología.
2. Basadas en usar la ontología en una aplicación y evaluar los resultados.
3. Aquellas que implican realizar comparaciones con una fuente de datos (e.g. una colección de documentos) sobre el dominio de interés.

4. Aquéllas en las que los humanos realizan la evaluación, tratando de valorar cómo de bien la ontología satisface criterios determinados.

Por otra parte, Brank y colegas también clasifican las aproximaciones en base a los elementos de la ontología en los que se centra la evaluación. En este caso, distinguen seis categorías o niveles:

1. Nivel léxico, de vocabulario o de datos.
2. Jerarquía o taxonomía (relaciones *is_a*)
3. Otras relaciones semánticas.
4. Nivel de contexto o aplicación.
5. Nivel sintáctico.
6. De arquitectura, estructura y diseño.

En su trabajo, Brank y colegas concluyen resaltando la necesidad de disponer de aproximaciones para la evaluación automática de ontologías, la cual consideran una precondición necesaria para el adecuado desarrollo de técnicas de procesamiento de ontologías para múltiples problemas (e.g. *ontology learning, population, matching, mediation, etc.*). En la tabla 3.1 se relacionan las dos clasificaciones proporcionadas por Brank y colegas.

Tabla 3.1. Aproximaciones para la evaluación de ontologías, relacionando los tipos de aproximaciones según el método de evaluación utilizado (horizontal), con el tipo de elementos de la ontología en los que se suele centrar cada aproximación (vertical). Adaptada de (Brank, et al., 2005).

Level	Approach to evaluation			
	Golden standard	Application-based	Data-driven	Assessment by humans
Lexical, vocabulary, concept, data	×	×	×	×
Hierarchy, taxonomy	×	×	×	×
Other semantic relations	×	×	×	×
Context, application		×		×
Syntactic	×			×
Structure, architecture, design				×

3.4.1 Evaluación para el desarrollo de ontologías

Esta sección está dedicada a aquellas técnicas ideadas para guiar el proceso de desarrollo o de “control de calidad” de una ontología. Se pretende minimizar el número de errores que contiene la ontología que se está desarrollando. Como afirma Gómez-Pérez (Gómez-Pérez, 2004), *“no es inteligente publicar una ontología o implementar una aplicación software basada en ontologías escritas por otros (o incluso por ti mismo) sin haber evaluado primero su contenido”*. Hace énfasis en que no existen artículos que describan el proceso que se debe seguir para evaluar ontologías antes de hacerlas públicas, ni herramientas adecuadas a lenguajes de representación de ontologías específicos, para facilitar esta evaluación.

Los primeros trabajos en evaluación de ontologías se centran en este tipo de evaluación (evaluación para el desarrollo) y, como ya se ha explicado, surgen alrededor de 1994, a raíz de la investigación de Asunción Gómez-Pérez y Thomas R. Gruber en el Laboratorio de Sistemas de Conocimiento (Knowledge Systems Laboratory, KSL)⁴³ de la Universidad de Stanford (Gómez-Pérez, 1994, 1995; Gómez-Pérez et al., 1995; Gruber & Olsen, 1994). En estos trabajos, se estudian los procesos de evaluación en el campo de la tecnología de compartición del conocimiento (Knowledge Sharing Technology, KST), que abarca tres elementos: (1) Las ontologías. (2) Los entornos software utilizados para desarrollarlas. (3) La documentación asociada a ellas. Gómez-Pérez (Gómez-Pérez, 1994) se muestra consciente de la carencia de mecanismos de evaluación en la KST y decide estudiar las similitudes y diferencias existentes entre bases de conocimiento y ontologías, tratando de aplicar ideas válidas para la evaluación de Sistemas Basados en Conocimiento (SBCs) a la KST.

El exhaustivo trabajo de revisión y comparación con los SBCs realizado, permite definir claramente los términos evaluación (*evaluation*), verificación (*verification*), validación (*validation*) y valoración (*assessment*) en el campo de la compartición del conocimiento. También se hace énfasis en la importancia de identificar “quién” se ocupará de la evaluación, pues esto determinará “qué” puede ser evaluado, “cuándo” se evaluará y “dónde” se desarrollará dicha actividad. Todo esto, teniendo en mente “por qué” es importante evaluar este tipo de tecnología. Así, explica que la evaluación

⁴³ <http://www-ksl.stanford.edu/>

puede ser realizada por el equipo de desarrollo de ontologías (evaluación) o por los usuarios finales (valoración o *assessment*). Mientras que el equipo de desarrollo de ontologías se ocupa de evaluar las propiedades técnicas de las ontologías, el software y la documentación, los usuarios finales valoran su utilidad y usabilidad. También resalta la necesidad de disponer de métodos, metodologías y herramientas que soporten el proceso de evaluación. Un año más tarde (Gómez-Pérez, 1995), describe un conjunto de ideas iniciales y generales que permiten guiar el proceso de evaluación de ontologías, y aplica dichas ideas a la evaluación de la ontología Bibliographic-Data (Gruber, 1994).

El trabajo realizado hasta ese momento por Gómez-Pérez en el campo de la evaluación de ontologías da como resultado, en 1996, un *framework* para la verificación de la KST (Gómez-Pérez, 1996) que se centra principalmente en la verificación de ontologías, pues el proceso de verificación de software para construir ontologías y de documentación ya lo proporcionaba el área de la ingeniería del software.

En aquella época, la ausencia de metodologías orientadas a dirigir el proceso de construcción de ontologías, provocaba que cada equipo de desarrollo aplicase sus propios principios, criterios de diseño y fases en el desarrollo de ontologías (Gómez-Pérez, 1999). Conscientes del problema de esta carencia metodológica, surgen varias aproximaciones ideadas para guiar el proceso de desarrollo de ontologías (Fernández et al., 1999; Gómez-Pérez & Rojas-Amaya, 1999; Gruninger & Fox, 1995; Uschold & Grüninger, 1996). Uno de los puntos comunes a estas metodologías es que todas contemplan la necesidad de evaluar las ontologías, lo cual contrasta con la falta de interés en aspectos de evaluación de la comunidad ontológica del momento, y llama la atención sobre la necesidad de dar mayor importancia a este proceso si se desea que las ontologías puedan proporcionar resultados satisfactorios en el ámbito empresarial. Aunque en aquel momento ya existían diversos métodos, técnicas y herramientas para evaluar Sistemas Basados en Conocimiento, éstos no se pueden reutilizar directamente para la evaluación de ontologías. Las ontologías se formalizan usando conceptos organizados en taxonomías, instancias, relaciones, funciones y axiomas, usando lenguajes formales. Estos elementos requieren de métodos de evaluación específicos.

Gómez-Pérez (Gómez-Pérez, 2004) propone un método para evaluar la consistencia, compleción y concisión de las taxonomías de ontologías (no tiene en cuenta los criterios de extensión ni de sensibilidad, también propuestos por la autora),

para su uso durante el desarrollo de ontologías. Para esto, presenta una clasificación (ver figura 3.5) de los posibles errores que pueden cometer los ontólogos al construir conocimiento taxonómico en una ontología, bajo una aproximación basada en *frames*.

Uno de los métodos más formales y explícitos para la evaluación de ontologías ideados hasta el momento es la metodología **OntoClean** (Guarino & Welty, 2002; Guarino & Welty, 2009), que se centra en “limpiar” la taxonomía de la ontología de relaciones subclase-superclase (i.e. *subclass-Of* o *is_a*) erróneas mediante un examen sistemático y riguroso de las meta-propiedades de los conceptos (identidad, unidad, dependencia y rigidez). OntoClean se aplica en dos pasos: en primer lugar, los conceptos se etiquetan de acuerdo a sus meta-propiedades; finalmente, se realiza una revisión de estas propiedades para detectar errores taxonómicos. En la figura 3.6 se puede observar la taxonomía de una ontología con varios errores. En la figura 3.7 se muestra la taxonomía corregida, tras aplicar el proceso de limpieza de OntoClean.

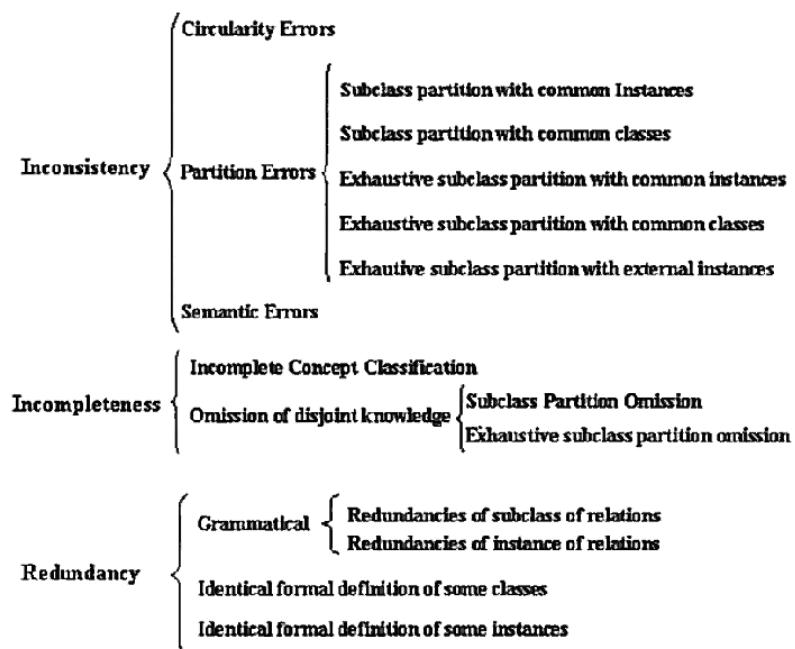


Figura 3.5. Posibles errores al desarrollar taxonomías de ontologías en *frames* (Gómez-Pérez, 2004).

La principal desventaja de OntoClean es que requiere un importante esfuerzo por parte de ingenieros de conocimiento expertos, familiarizados con las meta-propiedades de los conceptos. En general, esto ocurre con todos los métodos dedicados a evaluar las ontologías de acuerdo a los principios utilizados en su construcción. Como explican

Brewster y colegas (Brewster, et al., 2004), estos principios son válidos a nivel teórico, pero resulta extremadamente difícil construir pruebas automáticas que evalúen de forma comparativa varias ontologías en base a ellos. Esto es así porque se trata de principios de evaluación que dependen de semántica externa, que sólo los seres humanos son capaces de proporcionar. Debido a estos inconvenientes surge, en 2008, la herramienta **AEON** (Automatic Evaluation of ONtologies) (Völker et al., 2008), que es capaz de etiquetar conceptos con las meta-propiedades de OntoClean y realizar el chequeo de las mismas, todo ello de forma automática.

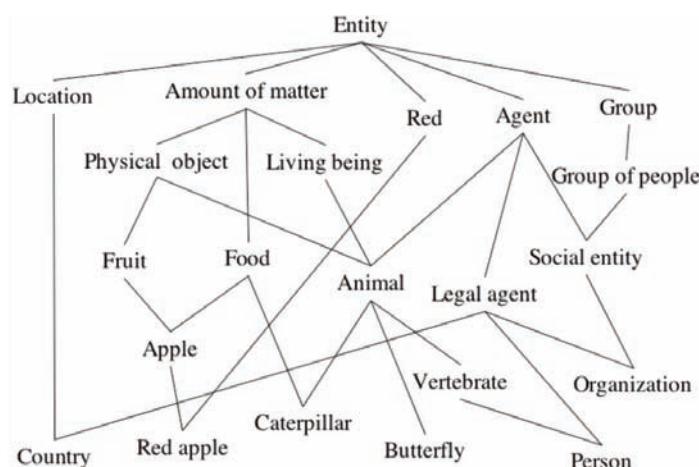


Figura 3.6. Ejemplo de taxonomía de una ontología antes de aplicar la metodología OntoClean. Las líneas representan relaciones *is_a* entre conceptos (Guarino & Welty, 2009).

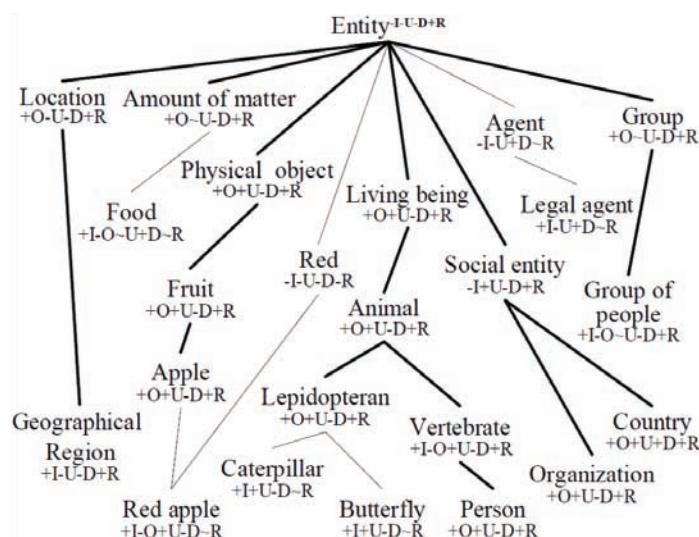


Figura 3.7. Taxonomía de la ontología tras aplicar OntoClean. Se puede observar que se han eliminado relaciones incorrectas (e.g. *Animal* es un *Physical Object*), o que se han creado nuevos conceptos (e.g. *Lepidopteran* como parente de *Caterpillar* y *Butterfly*). Las letras I, U, D y R se refieren a las metapropiedades de identidad, unidad, dependencia y rigidez, respectivamente (Guarino & Welty, 2009).

Otro de los trabajos ideados para detectar errores en ontologías es el de Köhler y colegas (Köhler et al., 2006), que proponen estrategias para identificar de forma automática términos y definiciones incorrectos en ontologías. En este trabajo se proporcionan métodos computacionales para medir la calidad de las definiciones de una ontología atendiendo a dos criterios: circularidad e inteligibilidad. La circularidad tiene que ver con si una definición usa el término que se pretende definir como parte de la definición (e.g. teléfono inalámbrico: teléfono sin cables), mientras que la inteligibilidad se refiere a evitar el lenguaje figurativo o difícil de entender. Los métodos propuestos se aplican a la conocida ontología Gene Ontology, permitiendo detectar más de 6000 términos problemáticos.

Otra medida que se puede utilizar para realizar un control de calidad de las ontologías durante su desarrollo es la propuesta por Buggenhout y Ceusters (Buggenhout & Ceusters, 2005). Estos autores proponen un método para calcular el “contenido de información” (*information content*) de los conceptos de ontologías, aplicable principalmente a ontologías de gran tamaño. Explican que la distribución de los valores proporcionados por esta medida en ontologías grandes puede proporcionar una impresión sobre la calidad de la ontología.

Teniendo en cuenta que una buena ontología es aquélla que proporciona buenos resultados en una tarea dada (Brank, et al., 2005), otro tipo de evaluación centrada en el desarrollo de la ontología consiste en observar los resultados obtenidos tras usar la ontología en una aplicación. Un ejemplo es el trabajo de Porzel y Malaka (Porzel & Malaka, 2004), en el que presentan una estrategia de evaluación que permite medir cómo de bien una ontología se adapta a una tarea determinada. Se trata de un esquema cuantitativo basado en la comparación con un *gold-standard* y dirigido a mejorar la calidad de la ontología. Sin embargo, este tipo de aproximaciones “basadas en la aplicación” presentan varios inconvenientes: (1) Se sabe si una ontología es buena o mala para una tarea después de usarla en dicha tarea, pero es difícil generalizar esta observación. (2) En casos en los que la ontología es un componente muy pequeño de la aplicación, sus efectos en los resultados pueden ser relativamente pequeños e indirectos. (3) Sólo es posible comparar varias ontologías si es posible conectarlas a la misma aplicación y observar los resultados (Brank, et al., 2005).

3.4.2 Evaluación para la selección de ontologías

En los trabajos citados en la sección anterior, se evalúan las ontologías desde la perspectiva de los principios utilizados en su construcción. Sin embargo, existe otro grupo de aproximaciones de evaluación de ontologías, posteriores, que surgen motivadas por la necesidad de reutilizar el conocimiento de ontologías ya existentes. Estas aproximaciones, en lugar de chequear si una ontología está bien construida o no, se centran en medir cómo de bien se ajusta a la resolución de un determinado problema. En este apartado, se presentan los principales ejemplos de este tipo de aproximaciones.

Existe un primer grupo de aproximaciones que permiten realizar una búsqueda básica de ontologías. Normalmente, estas aproximaciones exploran la Web Semántica en busca de ontologías y las indexan. Después, permiten a los usuarios realizar búsquedas por palabras clave y devuelven aquellas ontologías que contienen las palabras clave en alguno de sus nombres de concepto y, o, nombres de las propiedades de los conceptos de cada ontología. Como explican Jonquet y colegas (Jonquet, et al., 2010), esta funcionalidad, aunque útil para algunos casos, no es suficiente para poder considerar estas aproximaciones como aproximaciones de selección propiamente dichas, pues no disponen de la capacidad de decirle al usuario: “*éstas son las ontologías que deberías usar para tus datos*”. Sin embargo, resulta de interés conocer las principales aproximaciones en este sentido antes de presentar las aproximaciones de selección de ontologías, pues las aproximaciones de búsqueda pueden utilizarse como punto de partida para construir sistemas de selección.

Un ejemplo de aproximación de búsqueda de ontologías es **OntoKhoj** (Patel et al., 2003), que es un portal semántico que proporciona servicios de búsqueda, ordenación (ranking), agregación y clasificación de ontologías extraídas de la Web Semántica. Se basa en un algoritmo de ordenación de ontologías, denominado OntoRank, que está influenciado por el PageRank de Google (Page, et al., 1998). Dada cualquier ontología de la Web Semántica, este algoritmo proporciona un valor que indica su importancia o relevancia. Este valor se calcula en base a las referencias a la ontología desde y hacia otras ontologías, pero no todas las referencias entre ontologías se tratan igual: por ejemplo, si una ontología define una subclase de una clase de otra ontología, esta

referencia se considera de más valor que si la ontología sólo hace uso de una clase de otra ontología. Sin embargo, como explican Lewen y colegas, y Supekar (Lewen, et al., 2006; Supekar, 2005), esta forma de calcular la importancia de las ontologías no es adecuada ya que la tasa de reutilización de unas ontologías por otras es muy baja y, por lo tanto, las referencias entre ontologías son escasas. Esto desemboca en el “problema del huevo y la gallina”: los actuales métodos para encontrar las ontologías apropiadas no funcionan debido a la falta de enlaces entre ontologías, y los usuarios no enlazan unas ontologías con otras debido a la ausencia de métodos para encontrar las ontologías existentes. Otro de los sistemas que se basan en el algoritmo PageRank es el popular buscador de ontologías **Swoogle** (Ding, et al., 2004) (ver figura 3.8), que utiliza referencias entre diferentes ontologías para definir un grafo a partir del cual computa una puntuación para cada ontología.



Figura 3.8. Página principal del buscador de ontologías Swoogle.

Otros ejemplos de aproximaciones de búsqueda de ontologías son **OntoSearch** (Zhang et al., 2004), **OntoSearch2** (Pan et al., 2006) y **Watson** (d'Aquin, et al., 2007). Tanto OntoSearch2 como Watson proporcionan al usuario la capacidad de realizar consultas formales sobre su repositorio de ontologías, utilizando el lenguaje SPARQL.

A continuación, y tras citar las principales aproximaciones de búsqueda de ontologías, se presentan los ejemplos más importantes para la selección de ontologías.

(ONTO)²Agent (Arpírez, et al., 2000) es un agente que proporciona asistencia en la búsqueda de ontologías, ayudando a encontrar las ontologías que satisfacen parcial o totalmente las necesidades del usuario. Este agente basa su funcionalidad en una ontología de ontologías (denominada Reference Ontology), es decir, una ontología que clasifica múltiples ontologías existentes. En la figura 3.9 se muestra un esquema general del proceso de desarrollo que se ha seguido para construir la Reference Ontology, su interacción con (ONTO)²Agent y las principales funcionalidades que este último proporciona.

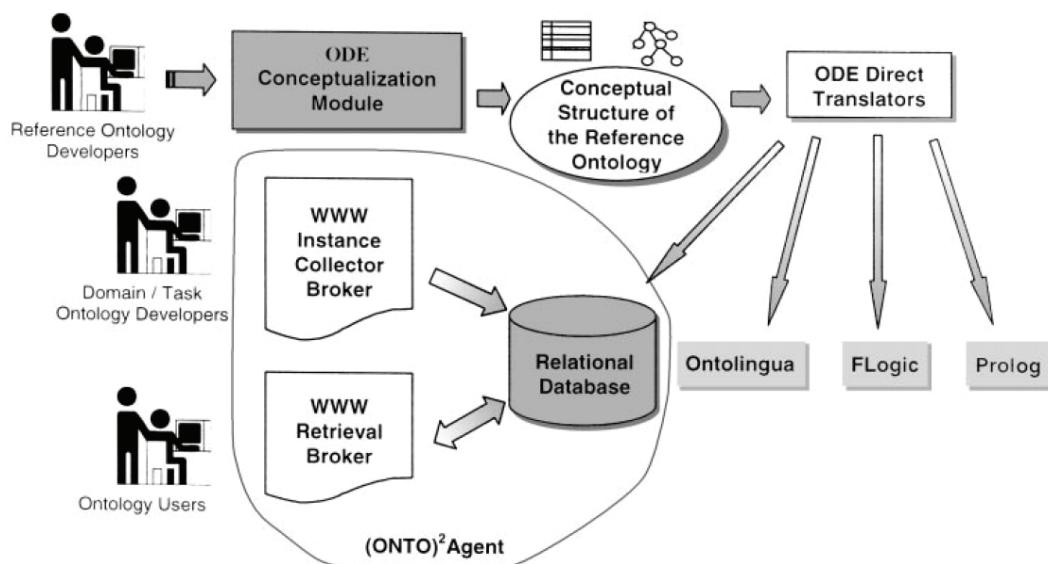


Figura 3.9. Visión general del proceso de desarrollo y funcionamiento de (ONTO)²Agent (Arpírez, et al., 2000).

Maedche y Staab (Maedche & Staab, 2002) indican que la tarea de buscar ontologías relevantes requiere medir la similitud entre ontologías de forma numérica (e.g. en el intervalo [0, 1]), y proponen varias medidas para medir la similitud entre ontologías a dos niveles: léxico y conceptual, de acuerdo a un *golden standard*. Se trata de un trabajo especialmente interesante en el campo del alineamiento de ontologías (*ontology alignment, ontology matching* (Euzenat & Shvaiko, 2007)), aunque también explican que a partir de esta idea pretenden desarrollar un buscador de ontologías que, en base a una ontología proporcionada por el usuario, recupere todas las ontologías similares a ella. El problema es que preparar el *golden standard* requiere mucho trabajo.

Gómez-Pérez también resalta la necesidad de disponer de métodos y herramientas para juzgar la usabilidad y utilidad de ontologías (ya desarrolladas y evaluadas) en situaciones concretas. Cuando los desarrolladores buscan ontologías candidatas para sus aplicaciones, se enfrentan a un complejo problema multi-criterio, y no elegir la ontología adecuada al problema concreto que se desea resolver provoca, en el futuro, que ésta se deje de usar y suele obligar a formalizar dicho conocimiento de nuevo (Gómez-Pérez, 2004).

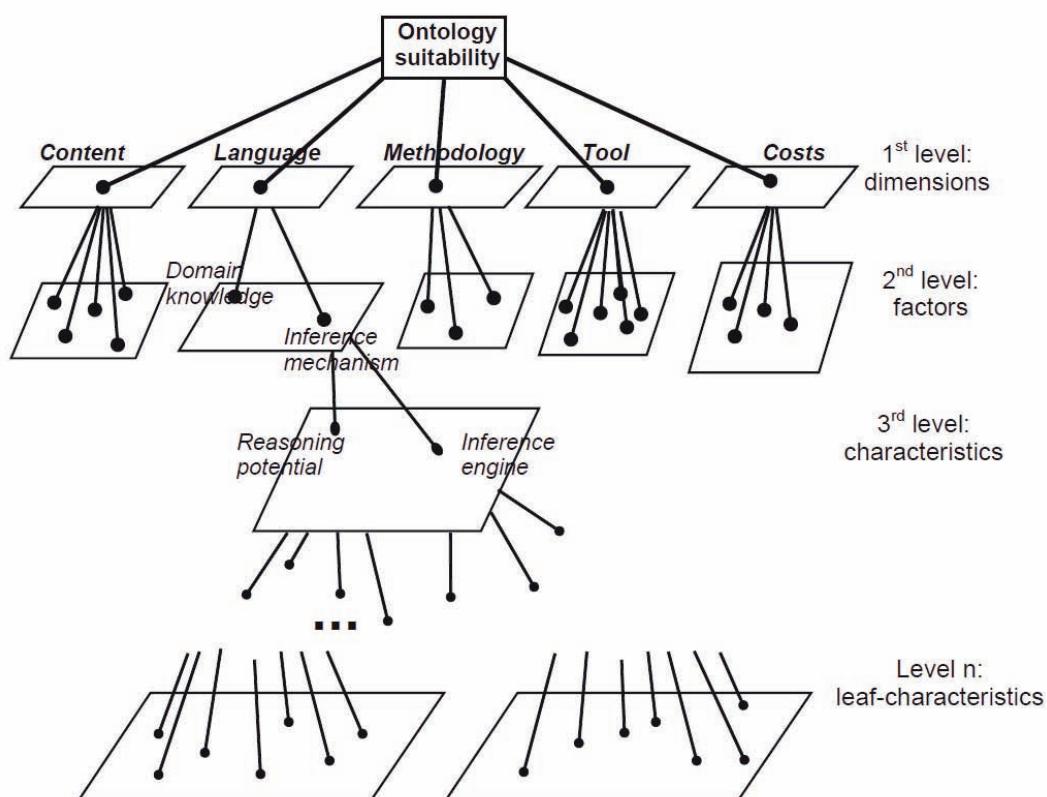


Figura 3.10. Representación del árbol de características (Lozano-Tello & Gómez-Pérez, 2004).

ONTOMETRIC (Lozano-Tello & Gómez-Pérez, 2004) es un método que permite a los usuarios medir la adecuación de ontologías ya existentes a una aplicación determinada. Lozano-Tello y Gómez-Pérez presentan una taxonomía de 160 características (llamada *multilevel framework of characteristics*), que proporciona el marco para seleccionar y comparar ontologías existentes (ver figura 3.10). Estas características constituyen la base para construir una ontología del dominio de las ontologías, conocida como la Reference Ontology. ONTOMETRIC basa su estrategias de evaluación en el proceso de jerarquía analítica (*analytic hierarchy process*) (Saaty, 1977), que

es un potente método para la toma de decisión en problemas multi-criterio complejos. Sin embargo, ONTOMETRIC tiene la desventaja de que requiere que se especifiquen detalladamente las características de las ontologías que se desea evaluar. Ésta es una tarea complicada y que consume mucho tiempo, además de ser una tarea en gran medida subjetiva.

En 2004, Brewster y colegas (Brewster, et al., 2004) resaltan la importancia de que las aproximaciones de evaluación de ontologías sean *corpus-driven* (o *data-driven*), y proponen una arquitectura y una aproximación probabilística para calcular cómo de bien una ontología cubre un dominio a partir de textos en lenguaje natural sobre dicho dominio. A partir de un conjunto de documentos que representan un dominio determinado, extraen un conjunto de términos o palabras clave que representan el dominio. Despues, calculan el solapamiento existente entre estos términos y los términos de una ontología (e.g. nombres de conceptos). De esta manera, son capaces de medir cómo de bien la ontología encaja con el dominio. También explican que la evaluación de ontologías no es comparable a las tareas de evaluación que se realizan en campos como la recuperación de información o el procesamiento de lenguaje natural, pues las nociones de *precision* y *recall* no son claramente aplicables a la evaluación de ontologías. En este caso, dependen de aspectos que pueden dar lugar a diferentes interpretaciones. Indican que en un escenario de Web Semántica, es probable que uno tenga que elegir la ontología más adecuada a una aplicación o dominio determinado, de entre un amplio abanico de ontologías existentes.

OntoSelect (Buitelaar, et al., 2004) (ver figura 3.11) es una herramienta que permite realizar búsquedas automáticas sobre un repositorio de más de 1500 ontologías⁴⁴ de diferentes dominios. Cuenta con un método para la selección de las ontologías más adecuadas a cada caso particular, que se basa en valorar cada ontología de acuerdo a los siguientes criterios:

- **Cobertura del dominio.** Mide en qué grado la ontología cubre los términos de entrada proporcionados por el usuario.

⁴⁴ A 17 de agosto de 2010, el repositorio de OntoSelect contiene 1.530 ontologías.

- **Estructura.** Mide la complejidad de la estructura de la ontología. Se basa en la idea de que, en general, las ontologías más avanzadas y trabajadas cuentan con estructuras más complejas y ricas.
- **Conexión.** Mide el grado de conexión de la ontología con otras ontologías. Se mide en base al número y calidad de las ontologías importadas y que importan la ontología.

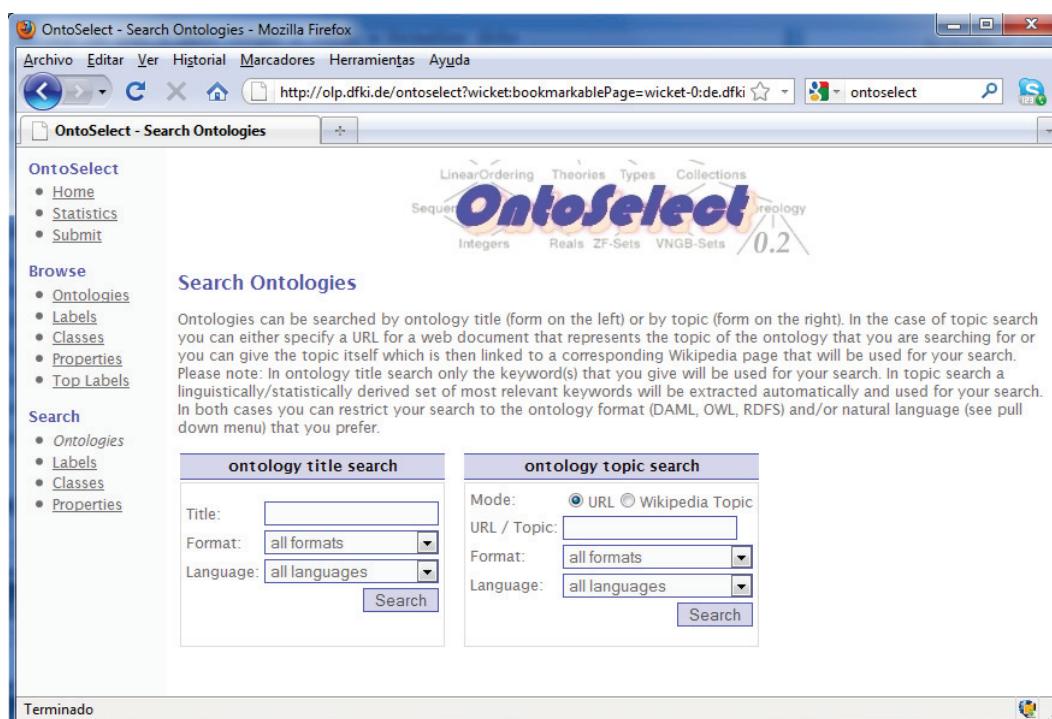


Figura 3.11. Menú de búsqueda de ontologías de OntoSelect (captura de <http://olp.dfki.de/ontoselect/>).

Sin embargo, estas tres medidas no se combinan en un único valor final que permita ordenar las ontologías de más a menos adecuadas al problema que se desea resolver. Éste es uno de los factores que provocan que las búsquedas de OntoSelect no proporcionen resultados adecuados para afrontar problemas reales.

Hong y colegas (Hong et al., 2005) proponen una aproximación interactiva no automática para seleccionar la ontología que mejor encaja con un dominio formalizado. Este método es similar a las aproximaciones de evaluación de ontologías basadas en comparar la ontología con un estándar (*gold standard*).

Otro trabajo que cabe citar en esta sección es el de (Supekar, 2005), que está dirigido a facilitar la creación y recuperación de metadatos de ontologías (e.g. dominio de la ontología, número de versión, autores, metodología utilizada, etc.). Viene motivado por la idea de que cuantos más datos se disponga de una ontología, más fácil y preciso resulta el proceso de evaluación. En su trabajo, se presenta una ontología de metadatos de ontologías (Metadata Ontology), que captura diversas características sobre ellas, y un sistema que usa esta ontología para permitir la creación y almacenamiento de metadatos de ontologías. Se pretende que un sistema de búsqueda de ontologías pueda realizar búsquedas sobre los metadatos y facilitar la elección de la ontología a utilizar.

Lewen y colegas (Lewen, et al., 2006) explican que reutilizar ontologías es esencial para alcanzar una buena interoperabilidad entre aplicaciones inteligentes, y para librarse a los ingenieros de conocimiento de la pesada tarea de construir ontologías de la nada pero que, sin embargo, existen dos razones por las cuales la reutilización de ontologías es poco frecuente: (1) Los actuales repositorios de ontologías proporcionan capacidades de búsqueda simples, basadas en una única palabra clave. (2) Incluso cuando un usuario encuentra una ontología que puede serle de ayuda, no dispone de información sobre su calidad y posibilidades de reutilización. Por esto, proponen un sistema para la valoración abierta de ontologías (denominado **Open Rating System** u ORS), que es capaz de proporcionar rankings de ontologías en base a: (1) Revisiones de la ontología como un todo. (2) Revisiones de dimensiones concretas de la ontología. (3) Valoraciones de estas revisiones. La idea básica de ORS es disponer de una aproximación democrática de valoraciones, donde cualquiera puede evaluar contenidos. Pero su poder real se basa en el concepto de *metarating*, que consiste en que los usuarios no sólo pueden valorar el contenido en sí mismo, sino también revisiones de contenido proporcionadas por otros usuarios (conocido botón del estilo “*¿Este comentario le ha resultado útil?*” de muchos sitios Web). Este sistema de valoración se ha implementado en **KnowledgeZone** (ver figura 3.12), un repositorio de ontologías Web al que los usuarios pueden enviar sus ontologías y anotarlas con metadatos. La información proporcionada por los usuarios se utiliza para elaborar rankings que resultan de utilidad para seleccionar la ontología más apropiada para una tarea. En este trabajo también se indica que, aunque es importante conocer la calidad de una

ontología como una función de aspectos cuantificables, es igual de importante conocer información subjetiva asociada a la ontología.

The screenshot shows the homepage of KnowledgeZone. At the top, there's a logo with a flower icon and the text "knowledge zone one stop shop for ontologies". Below the logo are links for "SEARCH & BROWSE", "SUBMIT ONTOLOGY", "STATISTICS", and "FAQ & NEWS". A search bar is present with the placeholder "search and browse ontologies". To the left, there's a sidebar with sections for "SEARCH", "BROWSE", and "SURVEY". The "SEARCH" section has a dropdown menu set to "All Ontologies". The "BROWSE" section lists categories like ALL ONTOLOGIES, ART ONTOLOGIES, BUSINESS ONTOLOGIES, COMPUTER ONTOLOGIES, GAME ONTOLOGIES, HEALTH ONTOLOGIES, HOME ONTOLOGIES, RECREATION ONTOLOGIES, REFERENCE ONTOLOGIES, REGIONAL ONTOLOGIES, SCIENCE ONTOLOGIES, SHOPPING ONTOLOGIES, SOCIETY ONTOLOGIES, and SPORTS ONTOLOGIES. The "SURVEY" section says "PLEASE TAKE OUR SURVEY". The main content area displays several ontology entries with details like name, added by, average rating, and a brief description. Some examples include "MGED Ontology", "User Modeling Ontology", "OBO relations ontology", "Foundational Model of Anatomy", and "TAMBIS". Each entry includes a "Read more" link.

Figura 3.12. Página principal de KnowledgeZone (Lewen, et al., 2006).

AKTiveRank (Alani, et al., 2006) es un prototipo de un sistema para la ordenación de ontologías que aplica varios métodos analíticos para valorar cada ontología en base a una estimación de cómo de bien representa un conjunto de términos de búsqueda dados. La arquitectura de AKTiveRank se muestra en la figura 3.13. En ella, se puede observar el funcionamiento del sistema, desde que el usuario realiza una consulta (paso 1) hasta que recibe una lista ordenada de ontologías relacionadas con la consulta (paso 7). El sistema se basa en el repositorio de ontologías de Swoogle y utiliza la librería JUNG (Java Universal Network/Graph) para trabajar con el grafo de cada ontología.

Para evaluar cada ontología, AKTiveRank se basa en cuatro métricas:

- *Class Match Measure* (CMM). Evalúa la cobertura de la ontología para los términos de búsqueda dados.
- *Density Measure* (DEM). Mide el detalle con el que se ha definido el concepto.
- *Semantic Similarity Measure* (SSM). Calcula la cercanía de los conceptos de interés en la estructura de la ontología. Se basa en la idea de que si los conceptos se encuentran próximos, es más fácil reutilizar una parte de la ontología.

- *Betweenness Measure (BEM)*. Calcula el número de caminos más cortos que pasan por cada nodo del grafo. Se basa en la idea de que si una clase tiene un valor alto para esta medida entonces es una clase central a la ontología.

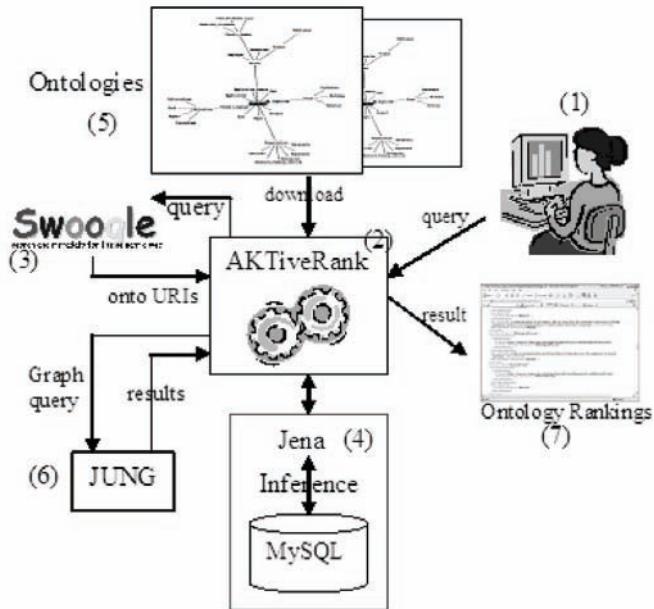


Figura 3.13. Arquitectura de AKTiveRank (Alani, et al., 2006).

Estas cuatro métricas se combinan en una única medida final (en el rango [0, 1]), que se utiliza para elaborar el ranking final.

Sabou y colegas (Sabou, et al., 2006a) proponen un algoritmo para la selección de ontologías en el contexto de herramientas que requieran reutilizar su salida de forma automática. En la figura 3.14 se pueden observar las principales tareas y etapas de este algoritmo.

Otro trabajo relevante es **WebCORE** (Cantador, et al., 2007). Se trata de un sistema que, a partir de una descripción informal de un dominio, determina qué ontologías de un repositorio son las más adecuadas para describir dicho dominio. Esta aproximación aplica varias técnicas de evaluación de ontologías de forma automática para proporcionar una lista ordenada de las ontologías más adecuadas. Para medir la popularidad de cada ontología, se basa en información introducida por los usuarios manualmente.

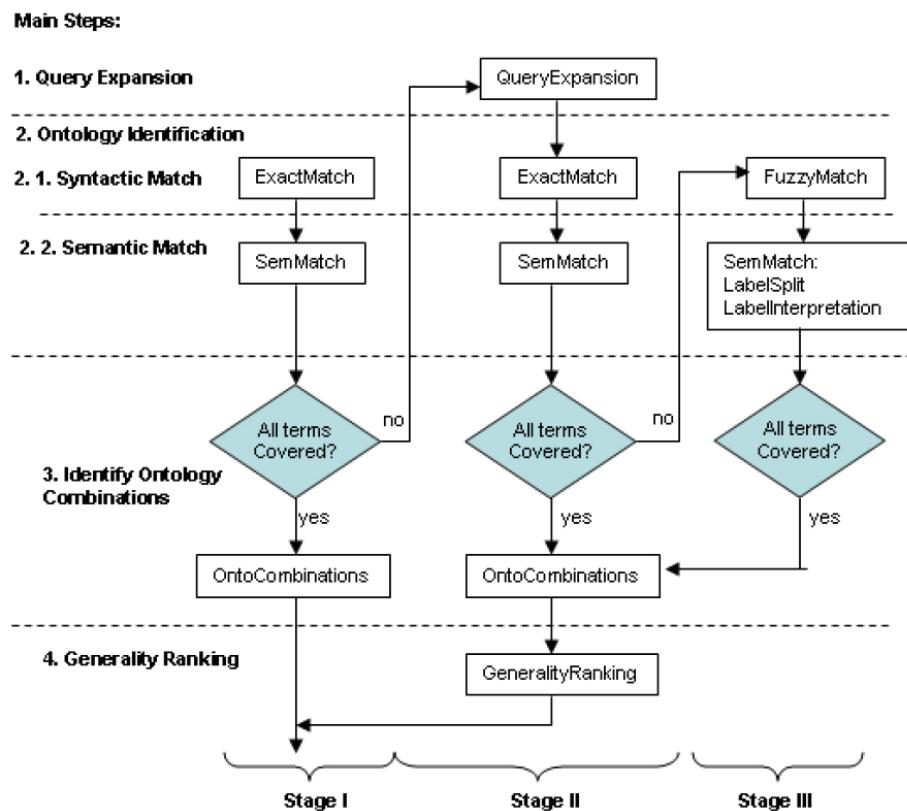


Figura 3.14. Principales tareas y fases del algoritmo de selección de ontologías (Sabou, et al., 2006a).

OntoQA (Tartir & Arpinar, 2007) es una herramienta que evalúa ontologías de acuerdo a varios criterios y proporciona como salida una lista ordenada de las ontologías que mejor encajan con un conjunto de palabras clave proporcionados como entrada. En esta aproximación, la evaluación tiene dos dimensiones: esquema de la ontología e instancias de la ontología. La primera dimensión evalúa el diseño de la ontología y su potencial para una representación rica de conocimiento. La segunda dimensión evalúa las instancias de la ontología. Como indican sus autores, OntoQA requiere poca implicación del usuario en el proceso de evaluación, aunque permite a los usuarios ajustar manualmente el proceso de ranking según sus necesidades particulares. La arquitectura de OntoQA se muestra en la figura 3.15.

DL-AOSF (Wang et al., 2008) (ver figura 3.16) es un *framework* para la selección automática de ontologías basado en Descripción Lógica, diseñado para varios escenarios de aplicación. El algoritmo se basa en dos criterios: cobertura del contexto y riqueza del conocimiento de la ontología. A diferencia de otras aproximaciones, esta

aproximación implementa los criterios de riqueza de conocimiento a un nivel semántico, utilizando razonamiento de lógica descriptiva.

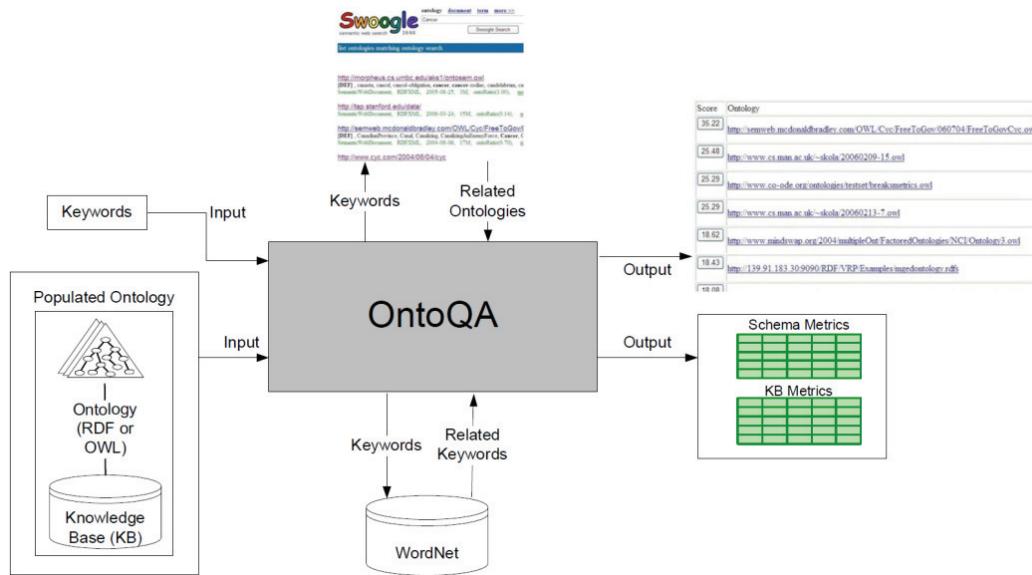


Figura 3.15. Arquitectura de OntoQA (Tartir & Arpinar, 2007).

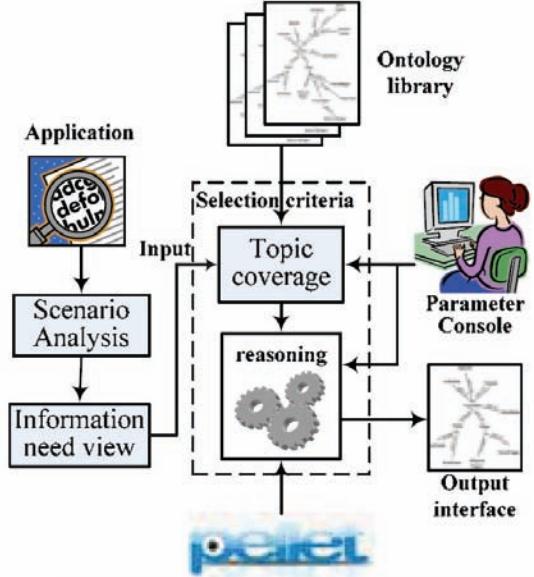


Figura 3.16. Arquitectura de DL-AOSF (Wang, et al., 2008).

CombiSQORE (Ungrangsi et al., 2008) propone un método que, a partir de una consulta semántica de entrada (descompuesta en subconsultas), realiza una expansión semántica usando WordNet y utiliza los resultados de motores de búsqueda de

ontologías como Swoogle y Watson para obtener las ontologías. Después, aplica varias métricas basadas en la similitud entre las subconsultas de entrada y las ontologías para proporcionar una lista ordenada que permita seleccionar la ontología más adecuada.

3.4.2.1 Aproximaciones aplicadas al ámbito de la biomedicina

Repositorios de ontologías biomédicas como el **NCBO BioPortal** (Whetzel, et al., 2009) y el **EBI Ontology Lookup Service** (Côté et al., 2006) proporcionan una funcionalidad de búsqueda que permite a un usuario consultar el repositorio en base a palabras clave. Estos repositorios mantienen un índice de nombres de conceptos y sinónimos para permitir realizar los tipos de búsqueda clásicos (correspondencia exacta o “contiene”). BioPortal también propone buscar en los valores de las propiedades de los conceptos permitiendo así, por ejemplo, realizar búsquedas basadas en el identificador del concepto. Ninguno de estos dos repositorios de ontologías ordena los resultados (e.g. como lo hace Swoogle). Se limitan a la búsqueda y, debido a esto, no se pueden considerar dentro de las aproximaciones de selección de ontologías.

En 2005, un trabajo de Natalya F. Noy y colegas (Noy et al., 2005), desde el NCBO, propone un nuevo enfoque en el campo de la evaluación de ontologías, en la línea de la Web 2.0. Ellos argumentan que uno de los aspectos esenciales para poder evaluar si una ontología que está pública en la Web es adecuada para una tarea particular o no, es permitir a los propios usuarios de las ontologías que escriban revisiones sobre ellas y las anoten. Indican que decidir si una ontología es buena o no, es una tarea subjetiva, ya que:

- Hay una gran carencia de medidas objetivas y computables para determinar la calidad de una ontología.
- Una ontología que es buena para una tarea puede no ser buena para resolver otra tarea.
- Aunque a la hora de decidir si una ontología es buena o no para una tarea, sería muy útil saber cómo la ontología se usó en el pasado y para qué aplicaciones resultó apropiada, ésta es una información que no suele encontrarse disponible.

En base a esto, afirman que disponer de revisiones y anotaciones sobre las ontologías, generadas tanto por los propios autores de la ontología como por sus

usuarios, es un componente crucial para hacer posible la reutilización de las ontologías, aunque en el trabajo citado se centran en la información proporcionada por los usuarios de las ontologías. Clasifican la información que se puede obtener de los usuarios de ontologías en 9 dimensiones: (1) Revisión y valoración general. (2) Información de uso. (3) Herramienta de verificación. (4) Cobertura del dominio. (5) Corrección. (6) Comentarios sobre conceptos específicos. (7) Correspondencias con otras ontologías y terminologías. (8) Ejemplos de lo que se puede hacer con la ontología. (9) Citas y referencias. Sin embargo, explican que permitir que cualquier persona, experta o no, pueda evaluar y valorar ontologías puede reducir la calidad de las evaluaciones. Como una solución a este problema, sugieren el uso de “Webs de usuarios de confianza”.

Este trabajo deriva en una aproximación para la evaluación de ontologías basada en la comunidad que actualmente se está integrando con BioPortal (Noy et al., 2009). BioPortal proporciona la infraestructura para permitir a usuarios proporcionar información como revisiones de ontologías para proyectos específicos o notas sobre evaluaciones de ontologías u opiniones sobre ellas. De esta manera, los propios usuarios proporcionan respuestas a preguntas como: ¿alguna vez ha usado alguien esta ontología de forma satisfactoria para una tarea similar a la mía?, ¿esta ontología está actualmente activa (i.e. continúa utilizándose y manteniéndose)? El NCBO está actualmente realizando experimentos con esta nueva idea, y todavía no se dispone de una aproximación de evaluación que la explote adecuadamente.

Sin embargo, y además de la búsqueda básica en BioPortal, el NCBO cuenta con un servicio para la recomendación de ontologías biomédicas⁴⁵ (**Biomedical Ontology Recommender Web Service**) (Jonquet, et al., 2010) que, a partir de un texto o un conjunto de palabras clave que describen el dominio, sugiere las ontologías más apropiadas para anotar o representar los datos. La evaluación de las ontologías se realiza de acuerdo a tres criterios:

1. Cobertura de los datos de entrada.
2. Conectividad. Las ontologías que contengan a los términos con más referencias desde otras ontologías, recibirán mayores puntuaciones.

⁴⁵ <http://stage.bioontology.org/recommender/>

3. Tamaño. Número de conceptos en la ontología.

En la figura 3.17 se puede ver un ejemplo de ejecución de sistema para un texto biomédico. Esta herramienta se puede considerar el trabajo de referencia actualmente para la selección de ontologías biomédicas.

The screenshot shows the NCBO Ontology Recommender Service interface. At the top, there's a search bar with the text "Basal cell carcinoma" and a search time of "3.0 s". Below the search bar are settings for "Recommendation scenario" (set to "Corpus"), "Normalize by ontology size" (unchecked), and "Repository" (set to "All Ontologies"). A text input field contains the query text. Below the input is a "Recommend" button and a "Clear" button. The main area displays a "Ontology Tag Cloud" with several terms in blue, including "Human developmental anatomy, timed ve", "SNOMED Clinical Terms, 2008_07_31", and "Clinical Terms Version 3 (CTV3) (Read Codes), 1999". A scroll bar indicates there are more terms below the visible area.

Figura 3.17. Servicio de recomendación de ontologías del NCBO.

Alani y colegas (Alani, et al., 2007) aplican una variante de su aproximación ya publicada (Alani, et al., 2006) al ámbito de la biomedicina. A partir del nombre de un dominio (e.g. anatomía), su sistema proporciona un conjunto de ontologías adecuadas para representar dicho dominio. Esta aproximación se basa en la idea de que las ontologías relevantes para un dominio o temática determinados, no suelen contener el nombre de dicho dominio en los nombres o propiedades de sus clases, ni en los valores de sus propiedades. Debido a esto, en este trabajo se realiza una expansión semántica usando información contenida en la Web para que, a partir del nombre del dominio, se pueda obtener un conjunto de palabras clave que lo representan (ver figura

3.18). Concretamente, se extraen estas palabras clave a partir de las páginas Web devueltas por un buscador Web al buscar por el nombre del dominio (en los experimentos han utilizado el buscador de Wikipedia, por proporcionar un mejor corpus para un dominio determinado que Google). A partir de estas palabras clave, se evalúan las ontologías de un repositorio para determinar cuáles son las ontologías que mejor las representan. En este trabajo no abordan el ranking de las ontologías seleccionadas.

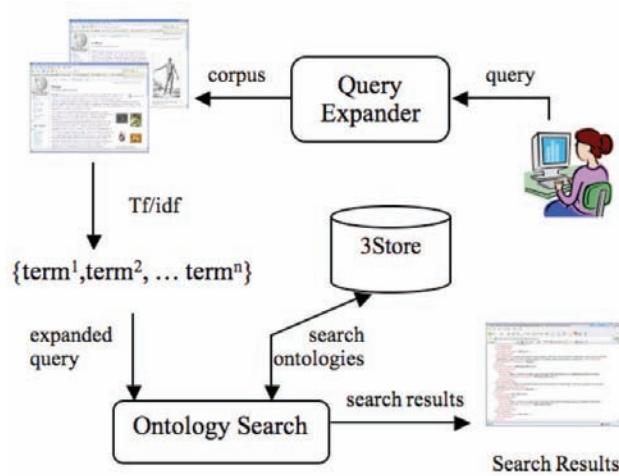


Figura 3.18. Proceso de búsqueda de ontologías con expansión de consulta.

Recientemente, Tan y Lambrix (Tan & Lambrix, 2009) proponen un *framework* para seleccionar la ontología más adecuada para una aplicación de minería de texto en el ámbito de la biomedicina. El *framework* está formado por tres componentes, cada uno de los cuales comprende varios criterios para realizar la evaluación. Se trata de un conjunto de criterios de alto nivel (ver apartado 3.2.1) que resultan útiles, y cualquier usuario debería tenerlos en cuenta a la hora de seleccionar una ontología en el ámbito de la minería de texto. Sin embargo, estos autores no proporcionan una forma automática de abstraer estos criterios a partir de una ontología y realizar una selección adecuada. Un usuario tendría que evaluar manualmente cada uno de ellos para decidir qué ontología reutilizar.

Otra aproximación para la selección de ontologías biomédicas es la de Maiga (Maiga, 2009). Se trata de una herramienta semi-automática que permite a un usuario especificar los requerimientos de una tarea en forma de una consulta que será

comparada contra un conjunto de ontologías. Una vez hecho esto, se utilizan varias métricas basadas en la coincidencia entre la consulta y el contenido de la ontología, así como en la proporción y densidad de las relaciones, para obtener un ranking de ontologías. En la figura 3.19 se puede observar el aspecto de la interfaz gráfica de esta herramienta.

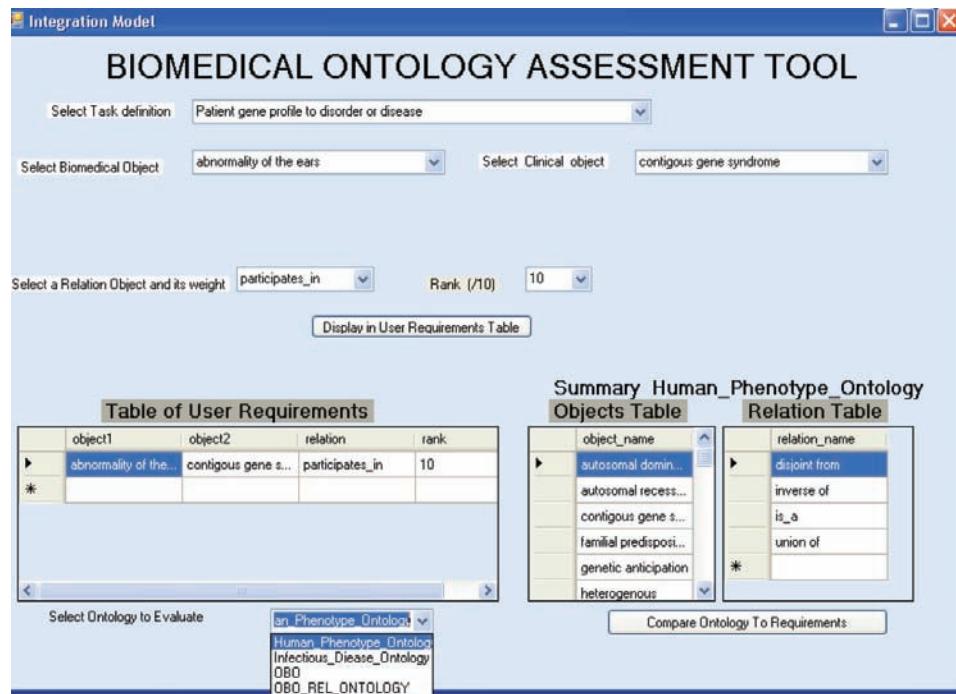


Figura 3.19. Interfaz gráfica de la herramienta para la selección de ontologías.

3.4.3 Comparativa de aproximaciones

En la tabla 3.2 se presenta una comparativa de las principales aproximaciones de selección de ontologías ideadas hasta el momento. Aunque siendo estrictos las aproximaciones que se limitan a la búsqueda no se considerarían en esta comparativa, éstas también se han incluido en la tabla, pues suponen la base de la selección de ontologías y se pueden reutilizar para construir sistemas de selección. La comparativa se basa en los criterios de evaluación propuestos por Jonquet y colegas (Jonquet, et al., 2010), ya explicados en el apartado 3.2.1. Se han añadido dos nuevos criterios: “metadatos” y “salida combinada”.

Tabla 3.2. Comparativa de aproximaciones para la selección de ontologías. Las celdas en verde indican un valor positivo para el criterio correspondiente, mientras que las celdas en naranja indican un valor negativo. Las celdas blancas indican indeterminado o no aplicable. Se resaltan las aproximaciones de selección del ámbito biomédico. Tabla adaptada y completada de (Jonquet, et al., 2010).

		Criterio												
		Aproximación												
Dominio biomédico	Búsqueda	Swoogle (Ding, et al., 2004)	Automatización	Dinamismo	Corresp. de términos	Corresp. propiedades	Expans. de consultas	Medidas estructurales	Conectividad	Desambiguación	Razonamiento	Popularidad	Metadata	Salida combinada
		Watson (d'Aquin, et al., 2007)												
Selección	Selección	OntoSearch (Zhang, et al., 2004)												
		OntoSearch2 (Pan, et al., 2006)												
		OntoKhoj (Patel, et al., 2003)	■		■	■	■	■	■	■				
		Búsqueda de BioPortal (Whetzel, et al., 2009)	■	■	■	■	■	■	■	■				
		EBI Ontology Lookup Service (Côté, et al., 2006)	■	■	■	■	■	■	■	■				
		Usuarios de BioPortal (Noy, et al., 2009)	■	■				■	■	■	■	■		
		Tan & Lambrix (Tan & Lambrix, 2009)			■	■								
		Maiga (Maiga, 2009)	■	■				■	■	■	■	■		
		Alani et al. (Alani, et al., 2007)	■		■	■	■	■						
		Recomendador del NCBO (Jonquet, et al., 2010)	■	■	■	■	■	■	■					
		Brewster et al. (Brewster, et al., 2004)			■	■	■							
		(ONTO) ² Agent (Arpírez, et al., 2000)	■	■		■	■					■		
		DL-AOSF (Wang, et al., 2008)	■		■		■		■	■	■			
		OntoSelect (Buitelaar, et al., 2004)	■		■	■	■	■	■	■				
		ONTOMETRIC (Lozano-Tello & Gómez-Pérez, 2004)		■								■		
		AKTiveRank (Alani, et al., 2006)	■		■			■	■	■				
		OntoQA (Tartir & Arpinar, 2007)		■	■									
		CombiSQORE (Ungrangsi, et al., 2008)		■			■	■						
		Hong et al. (Hong, et al., 2005)	■	■										
		WebCORE (Cantador, et al., 2007)	■	■	■	■	■			■	■			
		Sabou et al. (Sabou, et al., 2006a)							■	■				
		Supekar, 2005 (Supekar, 2005)	■	■								■		
		Lewen et al. (Lewen, et al., 2006)	■	■						■				

Metadatos. Se refiere a si el proceso de evaluación está principalmente basado en metadatos sobre la ontología (normalmente representados en una ontología, e.g. Metadata Ontology (Supekhar, 2005)), referentes al proceso de desarrollo o a características generales de la ontología (e.g. metodología o herramientas con las que fue desarrollada, lenguaje de desarrollo, dominio, etc.).

Salida combinada. Se refiere a si la aproximación proporciona como salida combinaciones de ontologías. Este criterio ha sido considerado como de gran importancia por varios autores (Sabou, et al., 2006b).

En la tabla 3.2, se puede observar que sólo existen 5 aproximaciones para la selección de ontologías en el ámbito biomédico. De ellas, sólo 2 son automáticas. Y de estas dos aproximaciones automáticas, ninguna contempla dos criterios que a día de hoy se consideran esenciales para una adecuada evaluación de ontologías: “popularidad” y “salida combinada”.

3.4.4 Limitaciones de las actuales aproximaciones

Observando la tabla comparativa presentada en el anterior apartado, y teniendo también en cuenta las limitaciones indicadas por Sabou y colegas (Sabou, et al., 2006b) y Jonquet y colegas (Jonquet, et al., 2010), a continuación se resumen los principales inconvenientes de las actuales aproximaciones para su aplicación a la selección de ontologías biomédicas.

- **No automáticas.** A pesar de que muchas de las aproximaciones existentes ya son completamente automáticas, muchas de ellas todavía delegan gran parte del trabajo de selección de ontologías en información solicitada de forma explícita a los usuarios. Esto supone una barrera importante para su futura integración en procesos de reutilización automática de conocimiento. Se dan estos casos:
 - Solicitan a los usuarios valoraciones sobre diferentes aspectos de la ontología y utilizan esta información para realizar la evaluación. Por ejemplo, el sistema WebCORE solicita a los usuarios valoraciones para medir la popularidad de las ontologías.

- Se basan en repositorios de ontologías que contienen metadatos adicionales, que deben ser introducidos manualmente por expertos o usuarios.
- **Sólo un término de entrada.** Aproximaciones como Swoogle y OntoKhoj únicamente contemplan búsquedas basadas en una palabra clave. Esta restricción supone una limitación importante para representar el contexto que se desea describir semánticamente. Como ya demostraron Sabou y colegas en 2006, la cobertura proporcionada por una aproximación de selección será mejor cuanto mayor sea el tamaño del conjunto de términos de entrada y su heterogeneidad (Sabou, et al., 2006a).
- **Formatos de entrada.** Actualmente, únicamente el Recomendador del NCBO es capaz de manejar ontologías en formatos diferentes de OWL y propios del ámbito biomédico, como el formato OBO o RRF (formato en el que se encuentra el Metatesauro UMLS).
- **Poco dinámicas.** Las herramientas no son lo suficientemente rápidas como para poder ser invocadas de forma dinámica por aplicaciones cliente, ni lo suficientemente precisas para evitar requerir de intervención humana.
- **No se tiene en cuenta la popularidad, o se mide de forma incorrecta.** A pesar de que la popularidad es un aspecto a tener en cuenta, únicamente 3 de las aproximaciones estudiadas tienen en cuenta este criterio. Y, en estos casos, se mide de forma inadecuada ya que, o bien se obliga a los usuarios a proporcionar información sobre las ontologías (e.g. WebCORE), o bien se evalúa la popularidad de acuerdo a los enlaces inter-ontología, lo cual varios autores han indicado que no es correcto (Sabou, et al., 2006b).
- **No proporcionan salida combinada.** Ninguna de las aproximaciones estudiadas es capaz de proporcionar combinaciones de ontologías a la salida. Varios autores han indicado que éste es un requisito indispensable para un sistema de evaluación de ontologías automático. A modo de ejemplo, en la figura 3.17 se puede ver que el Recomendador del NCBO, considerado el sistema de selección de ontologías biomédicas más avanzado hasta la fecha, proporciona una salida que consiste en una “nube” de los nombres de las

ontologías seleccionadas en la que los nombres de mayor tamaño representan las ontologías más recomendadas. Si un usuario necesita cubrir la mayor parte posible de su contexto, una salida como ésta no será suficiente, pues no proporciona una forma de ver cómo cubren el contexto diferentes combinaciones de las ontologías.

- **Se ignora la semántica proporcionada por la ontología.** Muchas de las actuales aproximaciones no tienen en cuenta la semántica proporcionada por los vínculos entre los conceptos de la ontología (relaciones *is-a*, *part-of*, etc.). De esta manera, se ignora información importante, como la distribución de los conceptos en la jerarquía de la ontología, los parientes de un concepto, conceptos que engloban a otros, etc. Como indican Sabou y colegas (Sabou, et al., 2006b), “*es sorprendente que ninguna de las actuales aproximaciones se beneficie del conocimiento ontológico, en lugar de tratar las ontologías como objetos interconectados, grafos o conjuntos de términos*”.
- **Se ignora el significado de los términos de entrada.** Como indican Jonquet y colegas (Jonquet, et al., 2010), en general, las aproximaciones no disponen de métodos de desambiguación, probablemente porque identificar correctamente los conceptos subyacentes a un conjunto de palabras clave de entrada es todavía una tarea muy compleja. Por ejemplo, si un usuario introduce como entrada a un sistema de selección el término *cold*, entre otros, es difícil averiguar si la intención del usuario es obtener resultados para la enfermedad (i.e. gripe) o para la sensación de frío. Esto se puede resolver teniendo en cuenta el contexto, es decir, los demás términos introducidos por el usuario. Sin embargo, la gran mayoría de las actuales aproximaciones todavía no disponen de esta funcionalidad.
- **Ausencia de herramientas.** La gran mayoría de aproximaciones existentes no cuentan con una implementación (e.g. una aplicación Web o un servicio Web) que se encuentre disponible para su uso por la comunidad biomédica.
- **Número de ontologías limitado.** Las aproximaciones trabajan en base a repositorios con un número muy bajo de ontologías biomédicas, lo cual limita su uso. La aproximación que contempla un mayor número de ontologías biomédicas es el Recomendador del NCBO, con unas 200 ontologías.

- **Selección de ontologías a evaluar.** Algunas aproximaciones no permiten seleccionar las ontologías a evaluar. La evaluación se realiza teniendo en cuenta siempre todas las ontologías existentes en el repositorio.
- **Métricas muy complejas.** Algunas de las aproximaciones existentes se basan en métricas muy complejas, que perjudican notablemente su rendimiento. Así, por ejemplo, AKTiveRank requiere de dos minutos para evaluar cada ontología (Sabou, et al., 2006b), lo cual es demasiado para tareas automáticas. Piénsese que, en este caso, completar una tarea de selección sobre un repositorio de 200 ontologías consumiría varias horas. Como indican Sabou y colegas (Sabou, et al., 2006b), “*es importante mantener un equilibrio entre la complejidad de los métodos de evaluación utilizados y el rendimiento de los algoritmos*”.
- **Uso de *mappings* entre ontologías.** No se hace uso de correspondencias entre ontologías. Sin embargo, recursos biomédicos como el UMLS Metathesaurus o BioPortal son muy ricos en *mappings*, que podrían utilizarse para mejorar el proceso de selección.

En esta tesis, y como se verá en los siguientes capítulos, se propone una aproximación para la selección de ontologías biomédicas que pretende superar muchos de estos inconvenientes.

4 Métodos

El trabajo de investigación que conforma esta tesis se ha llevado a cabo siguiendo las fases del método científico. En este capítulo, se explican los métodos que se han seguido, de cara a facilitar la comprensión del trabajo realizado y hacer posible su reproducción. De forma resumida, se han seguido los siguientes pasos:

En primer lugar, se ha analizado detenidamente el problema de la evaluación y selección de ontologías. Para esto, se han revisado los trabajos más relevantes en el campo, centrando la atención en aquéllos orientados al dominio biomédico. En el capítulo 1 de la tesis se puede encontrar una introducción al problema que se plantea.

Tras analizar el problema, se ha realizado un exhaustivo estudio de las aproximaciones existentes. Se ha revisado y aclarado la terminología utilizada en la evaluación y selección de ontologías, y se han analizado detenidamente las aproximaciones propuestas hasta la fecha en este campo, identificando sus principales limitaciones. Este trabajo se presenta en el capítulo 3.

Habiendo analizado el problema y las limitaciones de las aproximaciones existentes, se ha realizado un planteamiento del problema más detallado. Esto ha permitido fijar los objetivos del trabajo. Estos pasos se documentan en el capítulo 1.

Para solucionar el problema descrito, y teniendo en cuenta las carencias de las actuales soluciones al respecto, se ha planteado una nueva aproximación para la selección de ontologías biomédicas. Esta aproximación se fundamenta en una hipótesis, que se enuncia en el capítulo 1. En el capítulo 5, se presenta la aproximación elaborada en esta tesis. Se explica en qué consiste esta nueva aproximación, indicando qué es nuevo y en qué se diferencia de otras aproximaciones.

Tras plantear la aproximación, y con la finalidad de validarla, se ha construido un prototipo de sistema de selección de ontologías biomédicas basado en ella. Los detalles de construcción de este prototipo se describen en el capítulo 6.

Una vez finalizada la implementación del prototipo, éste se ha evaluado. Para esto, se ha diseñado un conjunto de casos de estudio representativos, cada uno de los cuales representa un escenario habitual de selección de ontologías. Con la ayuda de varios expertos en el ámbito de las ontologías biomédicas, se han ejecutado los casos de estudio y se ha estudiado el comportamiento de la aproximación propuesta para cada uno de ellos.

En último lugar, se han extraído las conclusiones del trabajo realizado, y se han planteado futuras líneas de investigación.

5 Aproximación propuesta

En este capítulo, se presenta la aproximación para la selección automática de ontologías biomédicas que constituye el núcleo de esta tesis doctoral. En primer lugar, y apoyándose en un diagrama de la aproximación, se describe de forma general el proceso de selección de ontologías y los principales componentes que conforman la aproximación. A continuación, se explica cada uno de estos aspectos de forma más detallada, justificando las decisiones adoptadas y haciendo uso de diversos ejemplos para facilitar la comprensión. Finalmente, se presentan las diferencias más importantes de la aproximación con las soluciones propuestas hasta la fecha.

5.1 Descripción general

En la figura 5.1, se muestra un diagrama de la aproximación que se propone en esta tesis. En él se pueden ver los principales componentes de la aproximación, así como observar los pasos que guían el proceso de selección de ontologías.

Se parte de un conjunto de términos biomédicos (términos de entrada), que representan el contexto que se desea describir semánticamente. El proceso de selección de ontologías se realiza en las siguientes fases:

1. **Expansión semántica.** La finalidad de esta fase es ampliar cada término de entrada con otros términos de igual significado (sinónimos). Estos sinónimos resultarán de utilidad durante el proceso de evaluación de ontologías, para incrementar las posibilidades de encontrar cada término en las ontologías candidatas.

2. **Recuperación de ontologías.** Consiste en obtener un conjunto de ontologías biomédicas, para su evaluación. La aproximación que se propone contempla la existencia de un repositorio o librería de ontologías biomédicas, en el que éstas se encuentran almacenadas.

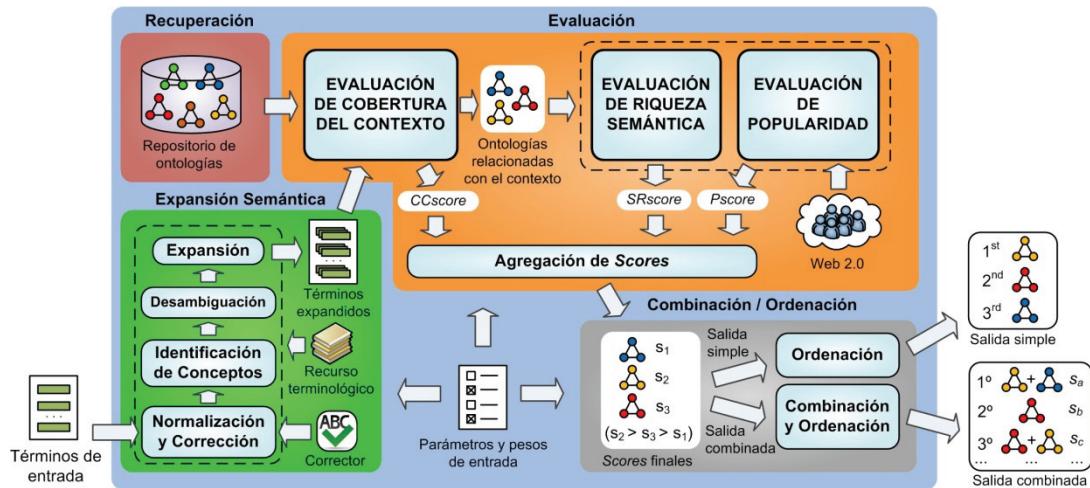


Figura 5.1. Proceso de selección de ontologías.

3. **Evaluación de ontologías.** Esta fase constituye el eje fundamental del proceso de selección. En ella, cada ontología se evalúa de acuerdo a tres criterios diferentes: (1) Cobertura del contexto, o en qué medida cubre cada ontología los términos de entrada. Las ontologías que no proporcionan al menos una cobertura mínima preestablecida (e.g. un 10% de los términos de entrada) se descartan, pues se asume que no resultarán de utilidad para describir los términos iniciales. (2) Riqueza semántica, o nivel de detalle que proporciona la ontología para cada término de entrada cubierto por ella. (3) Popularidad, o relevancia de la ontología en la comunidad biomédica. Para cada ontología, se obtiene una puntuación final (*score*) que indica su adecuación para describir semánticamente los términos iniciales.
4. **Combinación y ordenación.** La aproximación cuenta con dos tipos de salida: salida simple y salida combinada. La salida simple consiste en ordenar las ontologías de acuerdo a su puntuación final, obtenida en el paso anterior, dando lugar a un ranking en el que los puestos más altos los ocupan las ontologías más adecuadas para describir los términos iniciales. En la salida combinada, se calculan todas las posibles combinaciones de las ontologías

candidatas. Para cada combinación de ontologías, se obtiene una puntuación final combinada, calculada a partir de las puntuaciones individuales de cada ontología. Usando esta nueva puntuación, las combinaciones se ordenan y el ranking obtenido se proporciona como salida.

En los siguientes apartados, se explica de forma más detallada cada una de estas etapas.

5.2 Expansión semántica

La aproximación recibe como entrada un conjunto de palabras clave que representan un contexto determinado. Se asume que estas palabras clave estarán en inglés. Esta decisión se debe a que el idioma inglés es el más utilizado, con diferencia, en las ontologías biomédicas existentes actualmente.

El propósito de la fase de expansión semántica es ampliar cada palabra clave de entrada con otros términos que posean el mismo significado, o sinónimos. Estos sinónimos se utilizarán más adelante, durante el proceso de evaluación, para valorar en qué medida cubre los términos de entrada cada ontología candidata. El proceso de expansión semántica se realiza de acuerdo a los siguientes pasos:

1. **Normalización y corrección.** El objetivo de este paso es realizar una limpieza de los términos de entrada, transformándolos a un formato estándar, de tal manera que éstos puedan ser procesados de forma automática en los siguientes pasos. Los términos de entrada pueden contener errores tipográficos, que será necesario corregir, cuando sea posible, para asegurar una evaluación correcta. Además, si existen varios términos con el mismo significado, habrá que conservar únicamente uno de ellos y descartar los demás.
2. **Identificación de conceptos.** Una vez se dispone de los términos normalizados y corregidos, éstos se buscan en un recurso terminológico del ámbito biomédico. Existirán términos con un significado y términos con varios significados. También es posible que haya términos que no se encuentren en el recurso terminológico.

3. **Desambiguación.** En caso de que alguno de los términos de entrada tenga asociado más de un concepto (i.e. más de un significado), la finalidad de este paso es identificar cuál es el significado más apropiado para dicho término.
4. **Expansión.** Una vez se dispone de un concepto para cada término de entrada, y utilizando el recurso terminológico anteriormente mencionado, se genera el conjunto de términos equivalentes (sinónimos) de dicho término. Los conjuntos de términos obtenidos serán útiles a la hora de evaluar en qué medida cada ontología candidata cubre el conjunto de términos iniciales.

Como se puede ver en la figura 5.1, la expansión semántica hace uso de un recurso terminológico del ámbito biomédico, que proporciona el conocimiento del dominio necesario para realizar un adecuado tratamiento de los términos de entrada en el contexto de la biomedicina. Concretamente, el proceso de expansión semántica se basa en el Unified Medical Language System (UMLS), por ser el recurso terminológico biomédico más extenso y relevante actualmente. En apartado 2.1.5.2, se puede encontrar información detallada sobre UMLS.

A continuación, se describen de forma más detallada los pasos que conforman el proceso de expansión semántica.

5.2.1 Normalización y corrección

Las palabras clave que constituyen la entrada del sistema pueden contener números, abreviaturas, acentos, espacios en blanco, etc. Para poder procesar adecuadamente estos términos, es necesario convertirlos a un formato estándar, o lo que es lo mismo, normalizarlos. La normalización se realiza utilizando el recurso terminológico UMLS.

Además, los términos de entrada pueden contener errores ortográficos o tipográficos, que será necesario corregir cuando sea posible. También pueden existir términos de entrada duplicados, que habrá que descartar. La aproximación utiliza un corrector *online* (corrector de Yahoo⁴⁶) para realizar esta tarea.

⁴⁶ <http://developer.yahoo.com/search/web/V1/spellingSuggestion.html>

Ejemplo 5.1. Supóngase que se parte del siguiente conjunto de términos de entrada:

$$T = \{AORTA, Stomac, Pulmonar_Artery, diaprhamg, Cavity\}$$

Utilizando el recurso terminológico UMLS como base para realizar la normalización y el corrector ortográfico de Yahoo como corrector, los resultados del proceso de normalización y corrección son los que se muestran en la tabla 5.1.

Tabla 5.1. Ejemplo de normalización y corrección. Un “-” indica que no ha habido cambios.

Términos iniciales	Normalización	Corrección	Términos finales
AORTA	aorta	-	aorta
Stomac	stomac	stomach	stomach
Pulmonar_Artery	artery pulmonary	-	artery pulmonary
diaprhamg	-	diaphragm	diaphragm
Cavity.	cavity	-	cavity

Los términos obtenidos tras la normalización y la corrección están formateados adecuadamente para ser buscados en el recurso terminológico de referencia, e identificar así el significado o significados de cada uno de ellos. La fase de normalización y corrección está muy relacionada con la identificación de conceptos, que se explica a continuación.

5.2.2 Identificación de conceptos

La identificación de conceptos consiste en determinar el significado (o significados) de cada uno de los términos iniciales. Para esto, se buscan los términos de entrada, normalizados y corregidos (e.g. *aorta*, *stomach*, etc.), en el recurso terminológico de referencia. Al realizar esta búsqueda, pueden darse dos casos:

1. El término normalizado se ha encontrado en el recurso terminológico. En este caso, se obtienen y guardan los posibles significados (i.e. conceptos) asociados al término.
2. El término normalizado no se ha encontrado en el recurso terminológico. Esto puede deberse a que el término contenga algún error. Se usa un corrector ortográfico para corregir el término y se vuelve a buscar en el recurso terminológico. De nuevo, existen dos posibilidades:

- a. El término normalizado y corregido se ha encontrado. Se recuperan y guardan sus conceptos.
- b. El término normalizado y corregido no se ha encontrado, o bien no existe corrección para él. Esto puede deberse a dos causas:
 - i. El término es correcto, pero el recurso terminológico no lo contempla.
 - ii. El término es incorrecto, no existe. En este caso, debería ser descartado. Para averiguar si el término es correcto o incorrecto, se propone realizar una búsqueda del término en un buscador Web (e.g. Google, Yahoo!, etc.). Si el término supera un determinado número de ocurrencias, preestablecido (e.g. 10.000) en los resultados que proporciona el buscador, se considera que es correcto y se conserva. Si no, se considera que el término es erróneo y se descarta.

Como resultado del proceso de identificación de conceptos, se obtienen los conceptos en el recurso terminológico de referencia asociados a cada término inicial. En la figura 5.2, se muestra un diagrama del proceso de normalización, corrección e identificación de conceptos.

Ejemplo 5.2. Partiendo del conjunto de términos iniciales T del ejemplo 5.1, la tabla 5.2 muestra los conceptos identificados tras el proceso de identificación de conceptos, utilizando UMLS como recurso terminológico de referencia. Para cada concepto se muestra su identificador en UMLS (cui), su nombre más común, su tipo semántico (noción explicada en el apartado 2.1.5.2.2) y su definición.

5.2.3 Desambiguación

Para realizar una correcta expansión semántica, es necesario que cada término de entrada esté asociado a un único significado. En el caso de los términos para los que se han encontrado varios significados, habrá que identificar cuál de ellos es el más adecuado para el contexto particular, marcado por el conjunto global de términos de entrada. Este proceso se conoce como desambiguación terminológica (*word sense disambiguation*, WSD), y constituye un área de investigación en sí misma (Stevenson & Wilks, 2003).

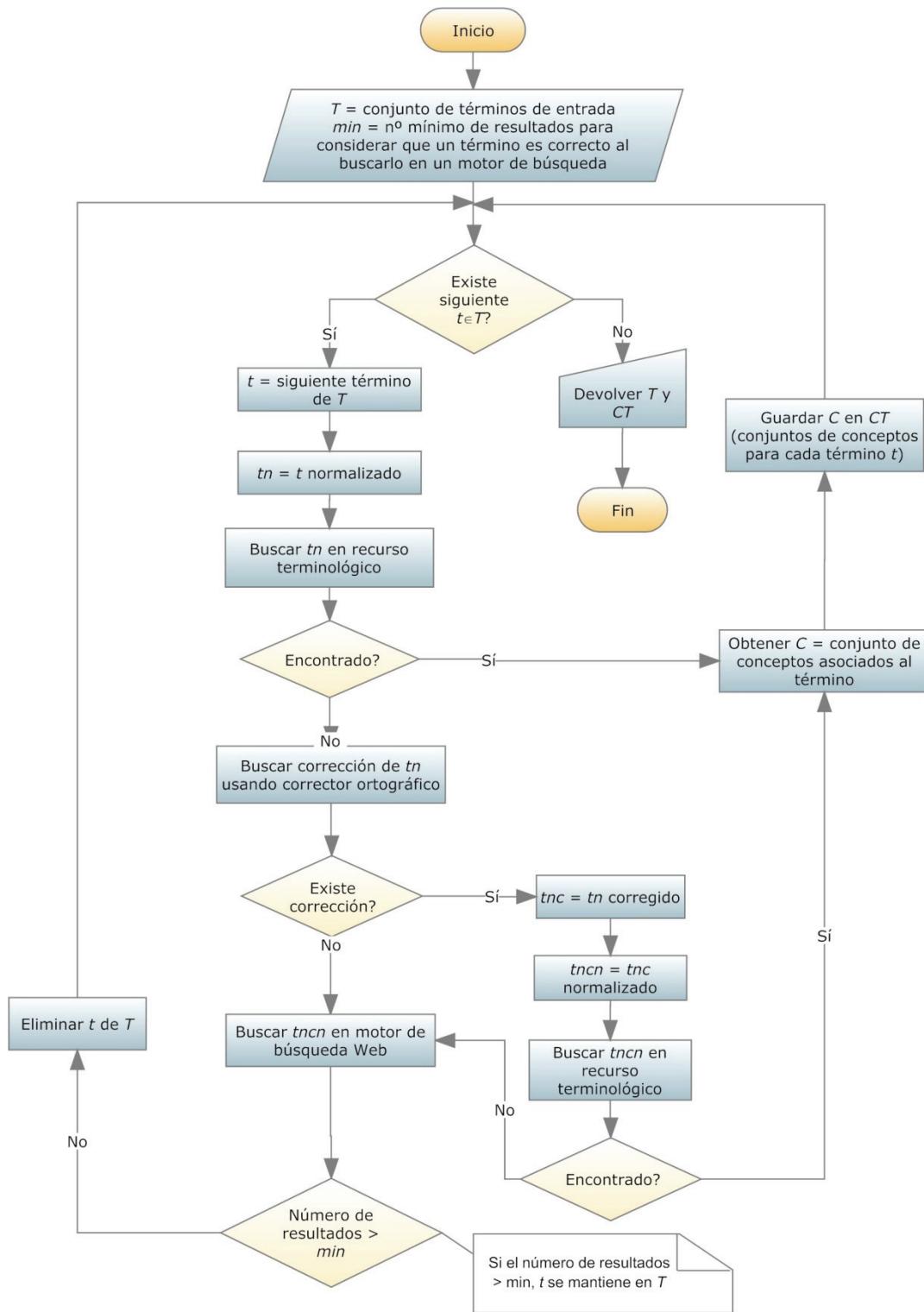


Figura 5.2. Proceso de normalización, corrección e identificación de conceptos.

Tabla 5.2. Ejemplo del proceso de normalización, corrección e identificación de conceptos. Para cada concepto identificado se muestra su identificador en UMLS (cui), su nombre más común, su tipo semántico y su definición.

Término inicial	Normalización y/o corrección	Conceptos identificados
AORTA	<i>aorta</i>	[C0003483: <i>aorta</i> (Body Part, Organ, or Organ Component)] The major arterial trunk that carries oxygenated blood from the left ventricle into the ascending aorta behind the heart, the aortic arch, through the thorax as the descending aorta and through the abdomen as the abdominal aorta (...)
<i>Stomac</i>	<i>stomach</i>	[C0038351: <i>stomach</i> (Body Part, Organ, or Organ Component)] An organ located under the diaphragm, between the liver and the spleen as well as between the esophagus and the small intestine. The stomach is the primary organ of food digestion.
		[C0038354: <i>stomach disorder</i> (Disease or Syndrome)] Condition in which there is a deviation from or interruption of the normal structure or function of the stomach.
<i>Pulmonar_Artery</i>	<i>artery</i> <i>pulmonary</i>	[C0034052: <i>pulmonary artery</i> (Body Part, Organ, or Organ Component)] The short wide vessel arising from the conus arteriosus of the right ventricle and conveying unaerated blood to the lungs.
<i>diaprhagm</i>	<i>diaphragm</i>	[C0011980: <i>diaphragm</i> (Body Part, Organ, or Organ Component)] The musculofibrous partition that separates the thoracic cavity from the abdominal cavity. Contraction of the diaphragm increases the volume of the thoracic cavity aiding inhalation (...)
		[C0042241: <i>diaphragm</i> (Medical Device)] A medical contraceptive device of soft flexible material, usually of thin rubber, that is designed to cover the cervix uteri prior to sexual intercourse to prevent the entry of spermatozoa.
		[C0152097: <i>diaphragm</i> (Disease or Syndrome)] Sin definición en UMLS.
		[C1705367: <i>diaphragm</i> (Medical Device)] Sin definición en UMLS.
		[C1705368: <i>vaginal diaphragm dosage form</i> (Biomedical or Dental Material)] A device usually dome-shaped, worn during copulation over the cervical mouth for prevention of conception or infection.
<i>Cavity.</i>	<i>cavity</i>	[C0011334: <i>tooth decay</i> (Disease or Syndrome)] Localized destruction of the tooth surface initiated by decalcification of the enamel followed by enzymatic lysis of organic structures and leading to cavity formation. If left unchecked, the cavity may penetrate the enamel and dentin and reach the pulp (...)
		[C0333343: <i>body cavity</i> (Body Space or Junction)] Compartment of trunk, which is embryologically derived from the intraembryonic celom; is located in the trunk, is enclosed by the body wall; contains serous sacs, viscera and other organs (...)
		[C1510420: <i>cavity</i> (Anatomical Abnormality)] Sin definición en UMLS.

El proceso de desambiguación utilizado se basa en la idea de Rindflesh y Aronson (Rindflesch & Aronson, 1994) de utilizar la Semantic Network de UMLS (ver apartado

2.1.5.2.2) como referencia para desambiguar. Esta noción consiste, en primer lugar, en determinar el tipo semántico preferido de acuerdo a un contexto. Una vez se cuenta con el tipo semántico preferido, se considera que los conceptos con el tipo semántico preferido son los que poseen el significado correcto.

Aunque la desambiguación que se propone se basa en la idea de Rindflesh y Aronson, los algoritmos para calcular el tipo semántico preferido y para determinar el mejor significado de un término en base al tipo semántico preferido, son una contribución propia. Previo a la explicación del proceso de desambiguación, resulta necesario definir la noción de frecuencia de un tipo semántico.

Definición 5.1 (frecuencia). Sea C un conjunto de conceptos de UMLS, y sea S el conjunto formado por los tipos semánticos de los conceptos de C (en el que puede haber elementos repetidos). Se define la *frecuencia de un tipo semántico* $s \in S$ como la función $fr: S \rightarrow \mathbb{N}$, que indica el número de repeticiones del tipo semántico s en S .

Teniendo en cuenta esta definición, el proceso de desambiguación que se propone se puede ver en la figura 5.3. A grandes rasgos, éste se divide en dos pasos:

1. **Identificación de tipo semántico preferido.** En base a los tipos semánticos de los conceptos encontrados en la etapa de identificación de conceptos y a la frecuencia de cada uno de ellos, se determina el tipo semántico preferido para los términos de entrada.
2. **Selección del mejor significado.** Usando el tipo semántico preferido, se determina el significado más adecuado para los términos ambiguos. Se considera que el significado más adecuado será aquél cuyo tipo semántico es el más cercano en la Semantic Network al tipo semántico preferido.

El método para calcular el tipo semántico preferido se muestra en la figura 5.6. El tipo semántico preferido para un conjunto de conceptos será aquel tipo semántico que mejor represente, de forma general, a dichos conceptos. En un primer momento, se podría pensar que el tipo semántico que puede actuar como mejor representante será aquél que sea más frecuente. Sin embargo, esto no tiene por qué ser así. Hay que considerar que la Semantic Network de UMLS no es una simple lista de tipos semánticos, sino una ontología en la que para cada tipo semántico se proporcionan varios tipos de relaciones con otros tipos semánticos. Explotar esta información

adecuadamente es crucial para lograr unos buenos resultados en el proceso de desambiguación.

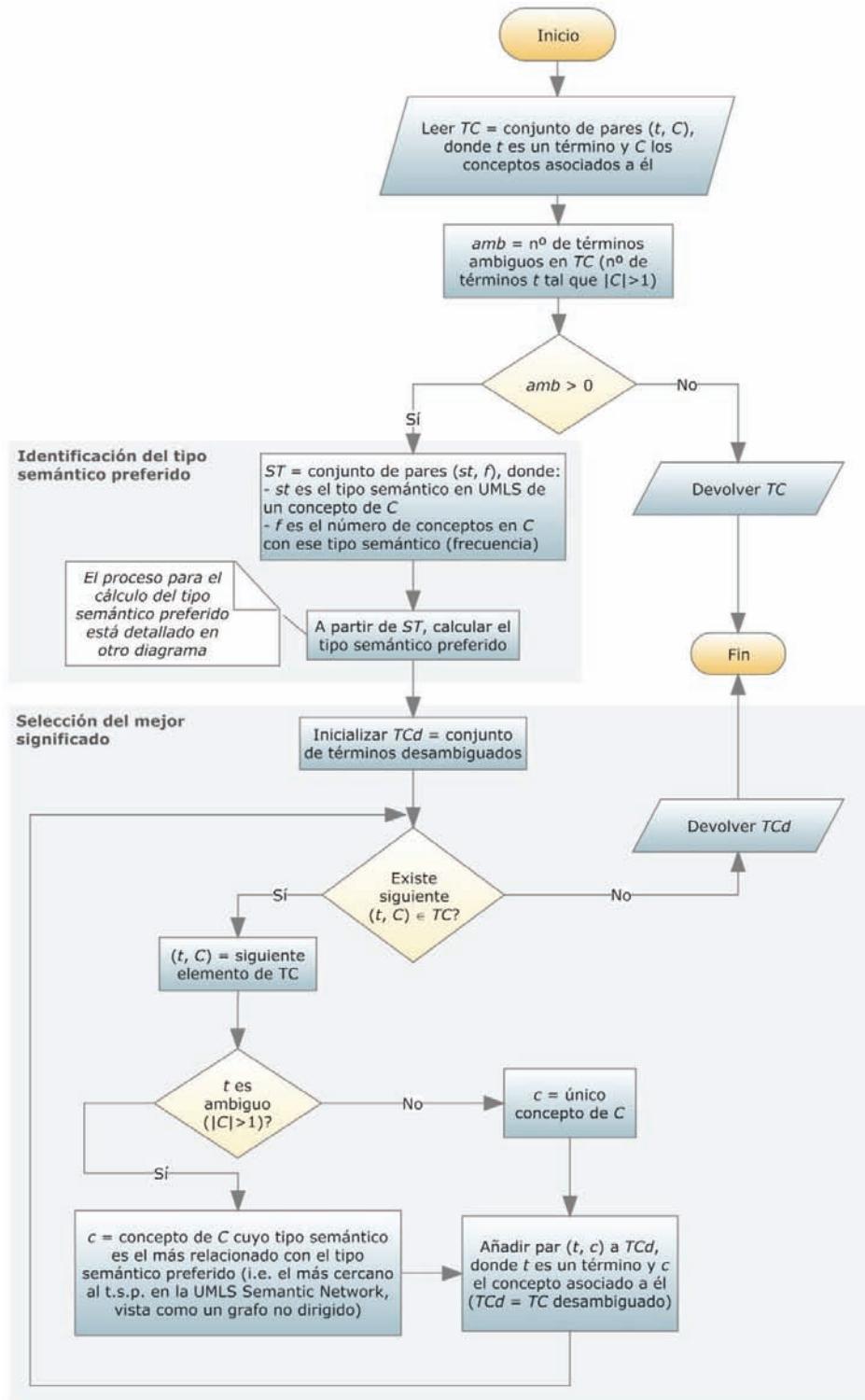


Figura 5.3. Proceso de desambiguación.

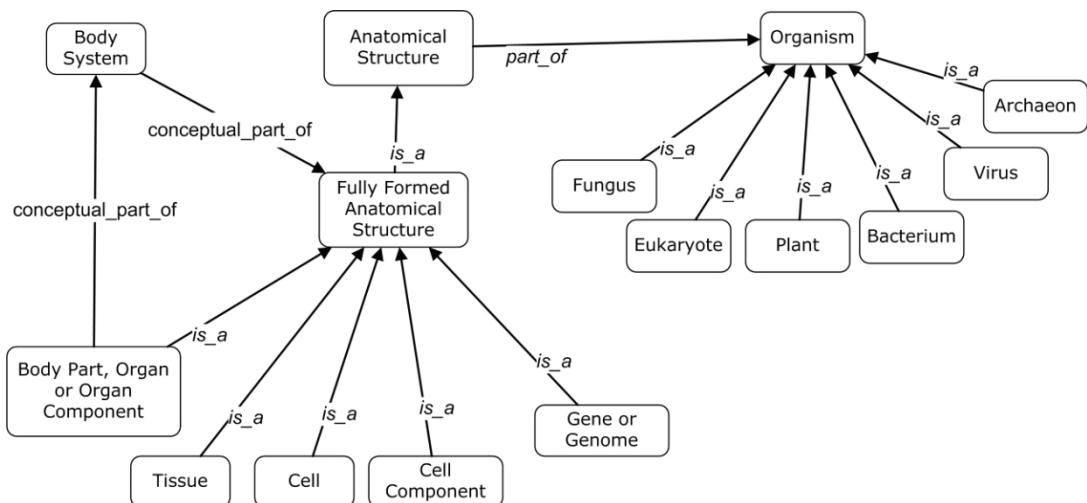


Figura 5.4. Fragmento de la Semantic Network de UMLS.

Ejemplo 5.3. Supóngase que los tipos semánticos para un determinado conjunto de conceptos, y sus frecuencias, son los que se muestran en la tabla 5.3.

Tabla 5.3. Ejemplo de tipos semánticos y sus frecuencias.

Tipo semántico	Frecuencia
<i>Cell</i>	1
<i>Tissue</i>	1
<i>Body System</i>	1
<i>Gene or Genome</i>	1
<i>Anatomical Structure</i>	1
<i>Fungus</i>	2

Observando únicamente la frecuencia de cada tipo semántico, podría parecer adecuado elegir el tipo semántico *Fungus* como tipo semántico preferido, pues su frecuencia es superior a la de todos los demás. Sin embargo, si se observa la estructura de la Semantic Network (ver figura 5.4), se puede ver que la mayoría de los tipos semánticos del ejemplo se agrupan a la izquierda de la figura, relacionados con el tipo semántico *Fully Formed Anatomical Structure*. Este tipo semántico sería un mejor representante de los demás tipos semánticos que el tipo semántico *Fungus*.

Este ejemplo da una idea de la importancia de tener en cuenta las relaciones de la Semantic Network al elegir el tipo semántico preferido, en el que se basará el proceso de desambiguación. Por ello, para calcular el tipo semántico preferido, en esta tesis se

propone la noción de tipo semántico más central o MCST (*Most Central Semantic Type*). Para explicar en qué consiste el MCST, es necesario introducir las nociones de distancia mínima y de distancia global.

Definición 5.2 (distancia mínima). Sea S un conjunto de tipos semánticos de la Semantic Network de UMLS, se define la *distancia mínima entre dos tipos semánticos* como la función $distmin: S \times S \rightarrow \mathbb{N}$, que proporciona la longitud del camino mínimo entre dos tipos semánticos cualesquiera $s_i, s_j \in S$, considerando la Semantic Network como un grafo no dirigido y no ponderado, en el que S es el conjunto de nodos y las aristas son las relaciones entre los tipos semánticos de S .

Definición 5.3 (distancia global). Sea S un conjunto de tipos semánticos de la Semantic Network de UMLS, con $|S| = n$. Sea $s \in S$ un tipo semántico cualquiera. La *distancia global de un tipo semántico* es una función $f: S \rightarrow \mathbb{N}$ definida como:

$$distglobal(s) = \frac{\sum_{i=0}^n distmin(s, s_i)}{fr(s)} \quad \forall s_i \in S$$

La distancia global es un indicativo de la centralidad del tipo semántico en el contexto que se contempla. Se puede observar que la suma de las distancias mínimas se divide por la frecuencia del tipo semántico, para favorecer la centralidad de los tipos semánticos que más se repiten. El MCST será aquel tipo semántico que minimice la distancia global. En la figura 5.5 se muestra el proceso de cálculo del conjunto de MCSTs.

Ejemplo 5.4. Continuando con los datos del ejemplo 5.2, los tipos semánticos para los términos iniciales, normalizados y corregidos, se muestran en la tabla 5.4. El número de ocurrencias, o frecuencia, de cada tipo semántico se muestran en la tabla 5.5. En este caso, el tipo semántico preferido es s_1 (*Body Part, Organ, or Organ Component*). Como se puede ver en la tabla 5.6, este tipo semántico es el que tiene la distancia global mínima.

Una vez calculado el tipo semántico preferido, el siguiente paso es realizar la desambiguación propiamente dicha, es decir, seleccionar el significado más adecuado para cada término. Se considera que el significado más adecuado será aquél cuyo tipo semántico posea la distancia mínima con el tipo semántico preferido. En la tabla 5.7, se muestra la distancia del camino más corto en la Semantic Network desde cada tipo

semántico al tipo semántico preferido. Estas distancias se utilizan para seleccionar el concepto que mejor representa cada término (ver tabla 5.8).

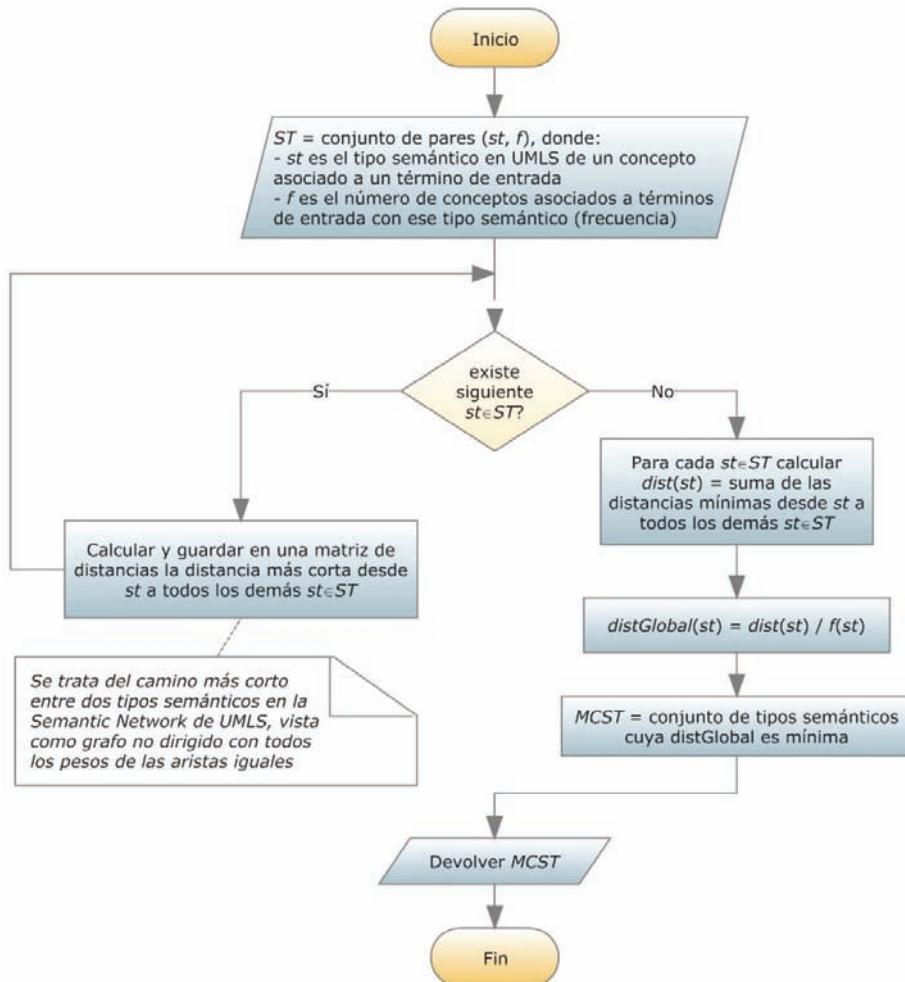


Figura 5.5. Cálculo del conjunto de tipos semánticos más centrales (*Most Central Semantic Types*, MCSTs).

En la tabla 5.8 se puede observar que para el término *cavity* existe un empate. Sus tres tipos semánticos poseen la misma distancia al tipo semántico preferido. En este caso, el empate se resolvería según se indica en la figura 5.6, calculando el tipo semántico más central entre ellos. En el caso de que se vuelva a producir otro empate, se realizaría el mismo proceso hasta obtener un único tipo semántico como resultado o hasta llegar a un punto en el que el algoritmo no avanza, obteniendo el mismo conjunto de tipos semánticos centrales una y otra vez. En este caso, se calcularía el tipo semántico más general (*lowest common ancestor*) entre ellos y éste sería el tipo semántico preferido.

Tabla 5.4. Tipos semánticos de los conceptos identificados para cada término.

Término	Conceptos identificados	Tipos semánticos
<i>aorta</i>	<i>aorta</i> (C0003483)	<i>Body Part, Organ, or Organ Component</i>
<i>stomach</i>	<i>stomach</i> (C0038351)	<i>Body Part, Organ, or Organ Component</i>
	<i>stomach disorder</i> (C0038354)	<i>Disease or Syndrome</i>
<i>artery</i> <i>pulmonary</i>	<i>pulmonary artery</i> (C0034052)	<i>Body Part, Organ, or Organ Component</i>
<i>diaphragm</i>	<i>diaphragm</i> (C0011980)	<i>Body Part, Organ, or Organ Component</i>
	<i>diaphragm</i> (C0042241)	<i>Medical Device</i>
	<i>diaphragm</i> (C0152097)	<i>Disease or Syndrome</i>
	<i>diaphragm</i> (C1705367)	<i>Medical Device</i>
	<i>vaginal diaphragm dosage form</i> (C1705368)	<i>Biomedical or Dental Material</i>
<i>cavity</i>	<i>tooth decay</i> (C0011334)	<i>Disease or Syndrome</i>
	<i>body cavity</i> (C0333343)	<i>Body Space or Junction</i>
	<i>cavity</i> (C1510420)	<i>Anatomical Abnormality</i>

Tabla 5.5. Tipos semánticos de los conceptos identificados y número de ocurrencias (frecuencia) de cada uno.

Tipo semántico	Frecuencia
<i>s₁: Body Part, Organ, or Organ Component</i>	4
<i>s₂: Disease or Syndrome</i>	3
<i>s₃: Anatomical Abnormality</i>	1
<i>s₄: Body Space or Junction</i>	1
<i>s₅: Biomedical or Dental Material</i>	1
<i>s₆: Medical Device</i>	2

Tabla 5.6. Distancias mínimas entre los tipos semánticos $s_i \in S$, suma de las distancias mínimas para cada tipo semántico ($\sum dmin$), frecuencias (fr), y distancias globales para cada tipo semántico ($dglob$).

	Distancia del camino más corto						$\sum dmin$	fr	$dglob$
	s_1	s_2	s_3	s_4	s_5	s_6			
s_1	0	2	2	2	4	2	12	4	3
s_2	2	0	2	1	4	3	12	3	4
s_3	2	2	0	1	3	1	9	1	9
s_4	2	1	1	0	4	2	10	1	10
s_5	4	4	3	4	0	4	19	1	19
s_6	2	3	1	2	4	0	12	2	6

Tabla 5.7. Distancia del camino más corto en la *Semantic Network* desde cada tipo semántico al tipo semántico preferido.

Tipo semántico	Distancia al t.s.p.
<i>s₁: Body Part, Organ, or Organ Component</i>	0
<i>s₂: Disease or Syndrome</i>	2
<i>s₃: Anatomical Abnormality</i>	2
<i>s₄: Body Space or Junction</i>	2
<i>s₅: Biomedical or Dental Material</i>	4
<i>s₆: Medical Device</i>	2

Tabla 5.8. Distancia mínima desde cada tipo semántico al tipo semántico preferido. Se resaltan las distancias mínimas para cada término. Se puede observar que en el caso del término *cavity* se produce un empate.

Término	Conceptos identificados	Tipos semánticos	Dist. al t.s.p.
<i>aorta</i>	<i>aorta</i> (C0003483)	<i>Body Part, Organ, or Organ Component</i>	0
<i>stomach</i>	<i>stomach</i> (C0038351)	<i>Body Part, Organ, or Organ Component</i>	0
	<i>stomach disorder</i> (C0038354)	<i>Disease or Syndrome</i>	2
<i>artery pulmonary</i>	<i>pulmonary artery</i> (C0034052)	<i>Body Part, Organ, or Organ Component</i>	0
<i>diaphragm</i>	<i>diaphragm</i> (C0011980)	<i>Body Part, Organ, or Organ Component</i>	0
	<i>diaphragm</i> (C0042241)	<i>Medical Device</i>	2
	<i>diaphragm</i> (C0152097)	<i>Disease or Syndrome</i>	2
	<i>diaphragm</i> (C1705367)	<i>Medical Device</i>	2
	<i>vaginal diaphragm dosage form</i> (C1705368)	<i>Biomedical or Dental Material</i>	4
<i>cavity</i>	<i>tooth decay</i> (C0011334)	<i>Disease or Syndrome</i>	2
	<i>body cavity</i> (C0333343)	<i>Body Space or Junction</i>	2
	<i>cavity</i> (C1510420)	<i>Anatomical Abnormality</i>	2

La tabla 5.9 muestra los resultados del proceso de desambiguación. Como se puede observar, cada término posee un único significado, o concepto, asociado. Este resultado se utilizará para ampliar el término de entrada con un conjunto de términos equivalentes a él, según se explica en el siguiente apartado.

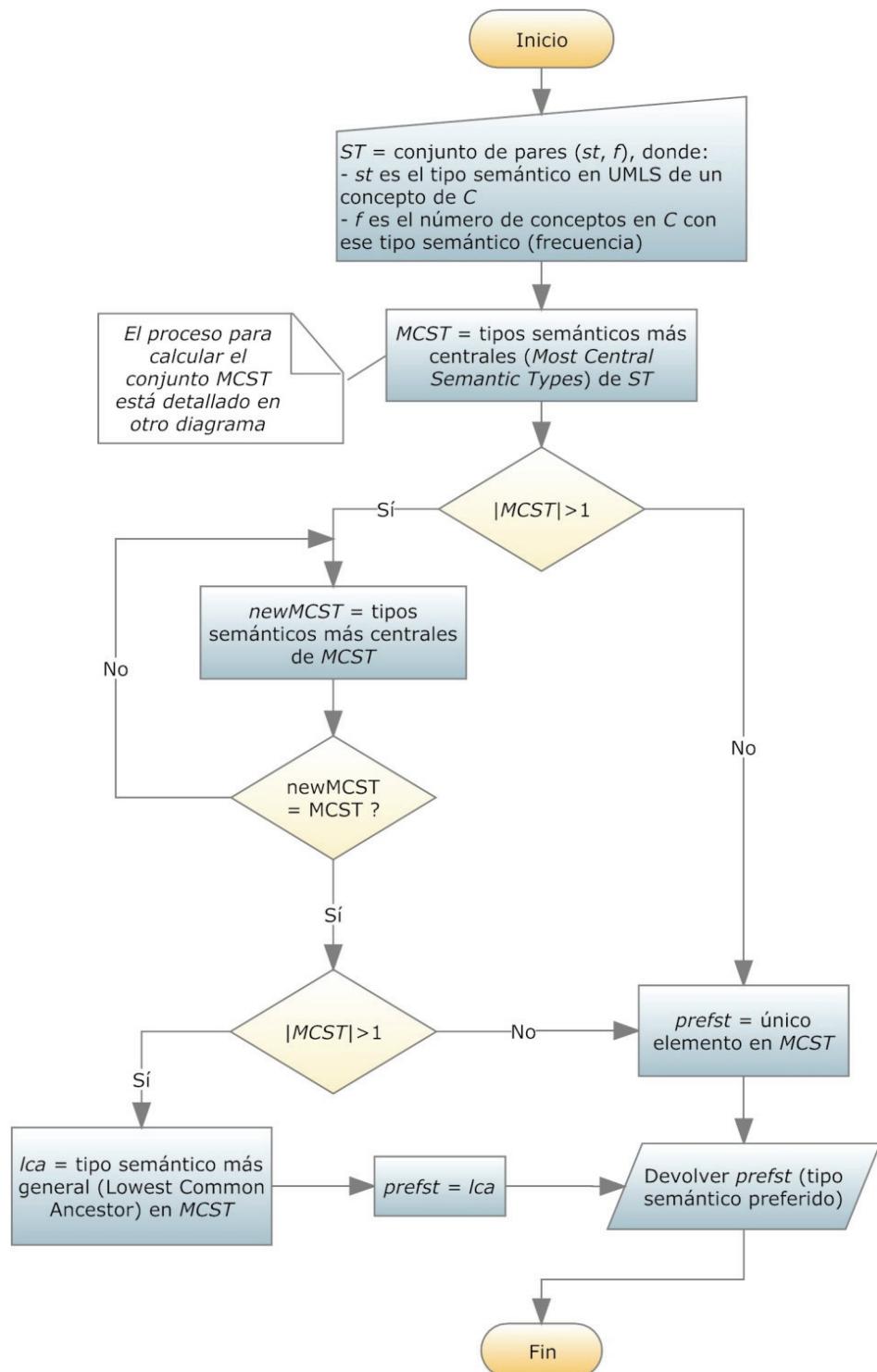


Figura 5.6. Proceso de cálculo del tipo semántico preferido.

Tabla 5.9. Concepto asociado a cada término de entrada una vez finalizado el proceso de desambiguación.

Término	Concepto resultado tras desambiguación
<i>aorta</i>	[C0003483: aorta (Body Part, Organ, or Organ Component)] The major arterial trunk that carries oxygenated blood from the left ventricle into the ascending aorta behind the heart, the aortic arch, through the thorax as the descending aorta and through the abdomen as the abdominal aorta (...)
<i>stomach</i>	[C0038351: stomach (Body Part, Organ, or Organ Component)] An organ located under the diaphragm, between the liver and the spleen as well as between the esophagus and the small intestine. The stomach is the primary organ of food digestion.
<i>artery pulmonary</i>	[C0034052: pulmonary artery (Body Part, Organ, or Organ Component)] The short wide vessel arising from the conus arteriosus of the right ventricle and conveying unaerated blood to the lungs.
<i>diaphragm</i>	[C0011980: diaphragm (Body Part, Organ, or Organ Component)] The musculofibrous partition that separates the thoracic cavity from the abdominal cavity. Contraction of the diaphragm increases the volume of the thoracic cavity aiding inhalation (...)
<i>cavity</i>	[C0333343: body cavity (Body Space or Junction)] Compartment of trunk, which is embryologically derived from the intraembryonic celom; is located in the trunk, is enclosed by the body wall; contains serous sacs, viscera and other organs (...)

5.2.4 Expansión semántica

Una vez se dispone de un único significado asociado a cada término de entrada, se realiza la expansión semántica del término. Esto consiste en acceder al recurso terminológico y obtener todos los términos equivalentes (sinónimos) a él. Los resultados del proceso de expansión semántica para los términos de la tabla 5.9 se muestran en la tabla 5.10.

Tras el proceso de expansión semántica, se dispone de varios conjuntos de sinónimos que se pueden considerar una representación semántica rica y libre de errores del contexto para el cual se está buscando una ontología. En base a esta información comienza el proceso de evaluación de ontologías, en el que se medirá numéricamente y de acuerdo a diferentes criterios, la adecuación de un conjunto de ontologías candidatas para describir el contexto. Los sinónimos de cada término resultarán esenciales para incrementar las posibilidades de encontrar el término en cada ontología, incrementando así la fiabilidad del proceso de evaluación.

Para llevar a cabo el proceso de evaluación de ontologías, es imprescindible disponer de una librería o repositorio de ontologías candidatas, para cada una de las cuales será necesario disponer de cierta información. Estos aspectos se tratan en el siguiente apartado.

Tabla 5.10. Resultado del proceso de expansión semántica.

Término	Concepto	Términos sinónimos
<i>aorta</i>	[C0003483: aorta (Body Part, Organ, or Organ Component)]	<i>aorta, aortas, aortic, trunk of systemic arterial tree, trunk of aortic tree</i>
<i>stomach</i>	[C0038351: stomach (Body Part, Organ, or Organ Component)]	<i>stomach, stomachs, gaster, gastr(o)-, ventriculus, gastric, gastrointestinal tract stomach</i>
<i>artery pulmonary</i>	[C0034052: pulmonary artery (Body Part, Organ, or Organ Component)]	<i>arteries pulmonary, artery pulmonary, pulmonary arteries, pulmonary artery, pulmonary trunk, truncus pulmonalis, artery pulm, pulm arteries, pulm artery, arteries pulm, pulmonary artery (trunk), pulmonary arterial trunk, trunk of pulmonary artery, pulmonary arterial tree organ part, pulmonary arterial tree, trunk of pulmonary arterial tree, pulmonary arterial tree (organ part), pulmonary artery structure</i>
<i>diaphragm</i>	[C0011980: diaphragm (Body Part, Organ, or Organ Component)]	<i>diaphragm, diaphragms, respiratory diaphragm, diaphragma, thoracic diaphragm, diaphragm respiratory, diaphragms respiratory, respiratory diaphragms</i>
<i>cavity</i>	[C0333343: body cavity (Body Space or Junction)]	<i>body cavity, cavity, cavities set, set of cavities, cavities, body cavities</i>

5.3 Recuperación de ontologías

La aproximación contempla la existencia de un repositorio de ontologías, en el que se almacenan todas las ontologías biomédicas candidatas durante el proceso de selección. La fase de recuperación de ontologías consiste en acceder al repositorio y obtener los datos de todas las ontologías que éste contiene, y cuya adecuación al contexto de entrada se pretende evaluar.

Existe un conjunto mínimo de información que este repositorio debe proporcionar, para satisfacer los requerimientos del proceso de evaluación de ontologías que se propone. Este conjunto mínimo se puede ver en la figura 5.7.

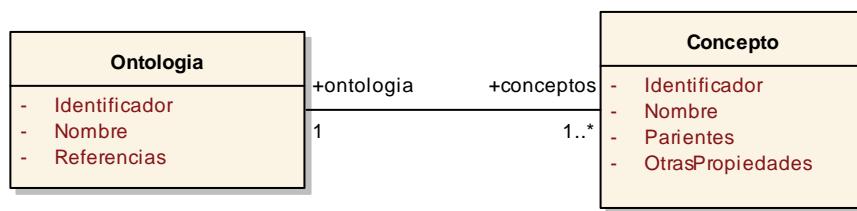


Figura 5.7. Modelo conceptual que muestra el conjunto mínimo de información que debe contener el repositorio de ontologías.

Esta información puede estar contenida explícitamente en el repositorio, o bien debe existir alguna forma de calcularla a partir de otra información almacenada en él. Por ejemplo, el número de parientes de un concepto podría estar almacenado en un campo del repositorio explícitamente, o se podría calcular a partir de la jerarquía de la ontología (relaciones *is_a* entre todos los conceptos de la ontología), para lo cual ésta tendría que encontrarse almacenada.

El repositorio guarda un conjunto de ontologías, cada una de las cuales debe contener como mínimo un concepto. Cada concepto almacenado en el repositorio pertenece a una única ontología. Para cada ontología, resulta necesario almacenar, al menos:

- **Identificador.** Un campo que identifique únicamente a la ontología en el repositorio (e.g. 36253).
- **Nombre.** Nombre de la ontología (e.g. Medical Subject Headings).
- **Referencias.** Número de referencias (o citas) a la ontología desde diferentes recursos Web, cuya información ha sido creada de forma colectiva por múltiples usuarios (e.g. Wikipedia, BioPortal, etc.). Esta información se utilizará para evaluar la popularidad o relevancia de cada ontología en la comunidad.

Mientras que para cada concepto de la ontología se debe disponer de:

- **Identificador.** Un campo que identifique a cada concepto del repositorio.
- **Nombre.** Término utilizado para representar al concepto (e.g. *white cell*). Esta información será necesaria para calcular la medida en que cada ontología cubre los términos de entrada.
- **Parientes.** Número conceptos de la ontología que son parientes directos del concepto según las relaciones taxonómicas (*is_a*). Es decir, se trata de

contabilizar el número total de padres, hijos y hermanos del concepto. Esta información se utilizará para evaluar la riqueza semántica de la ontología.

- **OtrasPropiedades.** Se refiere a la información adicional proporcionada para el concepto en la ontología. Se contabilizan las relaciones con otros conceptos (exceptuando las relaciones *is_a*), definiciones del concepto, sinónimos del nombre del concepto, etc. De la misma forma que el campo anterior, esta información se utilizará para medir el grado de riqueza semántica de la ontología.

Es necesario destacar que éste es un conjunto de datos mínimo para soportar la aproximación de selección planteada. En general, un repositorio de ontologías contendrá más información sobre cada ontología y sus conceptos (e.g. versión de la ontología, creador, ubicación del código fuente de la ontología, etc.), que resultará de gran utilidad a la hora de construir un sistema de selección de ontologías que resulte de utilidad.

Actualmente, se están realizando grandes esfuerzos para el desarrollo de repositorios de ontologías biomédicas a gran escala (e.g. iniciativa OBO Foundry⁴⁷, Workshop SERES 2010⁴⁸, etc.) y, probablemente, en varios años existan varios repositorios de ontologías de diferentes dominios, a los que se podrá acceder de forma rápida y sencilla. La aproximación aquí presentada no impone ningún límite respecto al carácter público o privado del repositorio, ni a su ubicación física. Únicamente es necesario que el repositorio disponga de la información mínima indicada anteriormente y que proporcione alguna forma de acceso a esta información.

En el siguiente capítulo de esta tesis (apartado 6.3), se explicará una posible implementación de un repositorio de ontologías de acuerdo a los requisitos anteriormente presentados, justificando adecuadamente las decisiones adoptadas.

⁴⁷ <http://www.obofoundry.org/>

⁴⁸ <http://www.ontologydynamics.org/od/index.php/seres2010/>

5.4 Evaluación de ontologías

El proceso de evaluación de ontologías constituye el núcleo del proceso de selección de ontologías y, por lo tanto, una parte crucial de la aproximación que se propone en este trabajo. Como se puede ver en la figura 5.1, este proceso consiste en evaluar cada ontología del repositorio de acuerdo a tres criterios diferentes, aunque complementarios:

1. **Evaluación de la cobertura del contexto.** Consiste en medir la cantidad de términos de entrada cuyo significado está contenido en la ontología.
2. **Evaluación de la riqueza semántica.** Consiste en valorar el nivel de detalle que la ontología proporciona sobre el contexto.
3. **Evaluación de la popularidad.** Tiene que ver con medir la relevancia de la ontología en la comunidad biomédica.

De cada tipo de evaluación, se obtiene un resultado numérico. Las tres puntuaciones obtenidas se combinan en un único valor, que indica la bondad de la ontología para describir el contexto.

A continuación, se explica en qué consiste cada tipo de evaluación y cómo se calcula la puntuación correspondiente. Finalmente, se explica cómo se realiza la agregación de las tres puntuaciones para cada ontología, en un único resultado.

5.4.1 Evaluación de la cobertura del contexto

La aproximación que se propone en este trabajo pretende seleccionar la mejor, o mejores ontologías para describir un determinado contexto, representado mediante un conjunto de términos o palabras clave. Debido a esto, uno de los principales aspectos a tener en cuenta al evaluar cada ontología candidata es el número de términos de entrada que se encuentran contenidos en ella. Aunque una ontología sea muy rica semánticamente o muy popular, no será adecuada para describir un conjunto de términos si éstos no están representados en la ontología.

Por ello, la finalidad de la **evaluación de la cobertura del contexto** es identificar las ontologías del repositorio que cubren el contexto. Se prefija un mínimo de

cobertura del contexto (e.g. 10%), de tal manera que aquellas ontologías que no aporten una cobertura superior a él serán descartadas y no se continuará su evaluación.

Para llevar a cabo la evaluación de la cobertura del contexto, se propone una medida llamada *Context Coverage Score (CCscore)*, que consiste en contar el número de términos de entrada (tras el proceso de normalización y corrección) que se encuentran contenidos en la ontología, considerando que un término está contenido en la ontología si existe una clase (concepto) en la ontología cuyo nombre se corresponde con el del término inicial, o con un término equivalente a él.

Para definir el *CCscore*, es necesario contar con una función que, a partir de un término, indique si el significado del término se encuentra en una ontología o no. Para esto, se definen antes las nociones de longitud de cadena, igualdad de cadenas, nombre de un concepto, nombres de los conceptos de una ontología, expansión semántica y *matching* exacto.

A partir de ahora, supóngase que \mathbb{S} es el conjunto de todas las posibles cadenas de caracteres (o palabras, o términos) sobre el alfabeto inglés, incluyendo además espacios y caracteres especiales.

Definición 5.4 (longitud de cadena). Sea s una cadena de caracteres perteneciente a \mathbb{S} . La *longitud* de s es el número de símbolos (incluyendo espacios en blanco) que componen la cadena. Se denotará mediante $|s|$. Por ejemplo, $|\text{skin cell}| = 9$.

Definición 5.5 (igualdad de cadenas). Sean dos cadenas de texto s y t , tal que $s, t \in \mathbb{S}$. Se dirá que s es *igual* a t , si tienen la misma longitud y los mismos símbolos en la misma posición. Se denotará mediante $s = t$. En caso contrario, se dirá que s es *distinto* de t , y se denotará como $s \neq t$.

Definición 5.6 (nombre de un concepto). Sea o una ontología, y C el conjunto de todos los conceptos contenidos en la ontología o , se define la función $\text{conceptname} : C \rightarrow \mathbb{S}$ como la función que proporciona el nombre de un concepto cualquiera $c \in C$.

Definición 5.7 (nombres de conceptos). Sea O un conjunto de ontologías, sea o una ontología de O , sea C el conjunto de todos los conceptos contenidos en la ontología o , y sea $\mathcal{P}(\mathbb{S})$ el conjunto potencia (o partes) de \mathbb{S} . Es decir, $\mathcal{P}(\mathbb{S})$ es el

conjunto de todos los subconjuntos de posibles cadenas de caracteres. Se define la función $\text{conceptnames} : O \rightarrow \mathcal{P}(\mathbb{S})$ como la función:

$$\text{conceptnames}(o) = \{s \mid s \in \mathbb{S} \wedge s = \text{conceptname}(c) \forall c \in C\}$$

Es decir, es la función que permite obtener el conjunto de todos los nombres de los conceptos en la ontología o .

Definición 5.8 (expansión semántica). Sea $s \in \mathbb{S}$ un término cualquiera, se define la función $\text{expand} : \mathbb{S} \rightarrow \mathcal{P}(\mathbb{S})$ como aquella función que proporciona los sinónimos de un término s dado, incluyendo al propio término.

Definición 5.9 (matching exacto). Sea $T \subseteq \mathbb{S}$ un conjunto de términos, y sea t un término de T . Sea O un conjunto de ontologías, y o una ontología de O . Se define la función $\text{matching} : T \times O \rightarrow \{0, 1\}$ de la forma siguiente:

$$\text{matching}(t, o) = \begin{cases} 1, & \text{si } \exists s \in \mathbb{S} \mid s \in \text{expand}(t) \wedge s \in \text{conceptnames}(o) \\ 0, & \text{en otro caso} \end{cases}$$

Es decir, la función toma el valor 1 si el término t o cualquiera de sus términos sinónimos coincide exactamente con el nombre de algún concepto de la ontología o . En caso contrario, su valor es 0. Si $\text{matching}(t, o) = 1$, se dirá que la ontología o cubre el término t , o que existe un *matching* entre el término t y la ontología o .

Ejemplo 5.5. Sea un término $t \in \mathbb{S}$ tal que $t = \text{medicine}$ y $\text{expand}(t) = \{\text{medicine}, \text{drug}, \text{remedy}, \text{medicament}, \text{medication}\}$. Sea una ontología o con cuatro conceptos, tal que $\text{conceptnames}(o) = \{\text{cell}, \text{heart}, \text{drug}, \text{blood}\}$. En este caso, $\text{matching}(t, o) = 1$ porque existe un término (el término *drug*), que pertenece al conjunto $\text{expand}(t)$ y también al conjunto de nombres de conceptos de la ontología o . En este caso, se diría que la ontología o cubre el término $t = \text{medicine}$, o que existe un *matching* entre el término t y la ontología o .

A partir de la función de *matching*, se define el *CCscore* de la siguiente manera:

Definición 5.10 (CCscore). Sea $T \subseteq \mathbb{S}$ el conjunto de términos resultantes tras el proceso de normalización y corrección de los términos de entrada, y sea O un conjunto de ontologías candidatas. Se define la puntuación de cobertura del contexto o *CCscore* (*Context Coverage score*), como una función $\text{CCscore} : O \times T \rightarrow [0, 1]$ tal que:

$$\text{CCscore}(o, T) = \frac{\sum_{i=1}^n \text{matching}(t_i, o)}{n} \quad \forall t_i \in T, \text{y con } n = |T|$$

Esta función calcula un valor en el intervalo $[0, 1]$ que indica la proporción de términos cuyo significado se ha encontrado en la ontología dada. Un 0 indicará que la ontología no contiene ninguno de los términos, ni ninguno de sus respectivos sinónimos. Un $CCscore$ de 1 indicará que la ontología evaluada cubre todos los términos de T , o lo que es lo mismo, que proporciona una cobertura del contexto de un 100%.

A continuación, se proporciona un ejemplo de cálculo del $CCscore$ para un conjunto de términos y una ontología determinados. En la figura 5.8 se puede ver una representación gráfica de dicho ejemplo.

Ejemplo 5.6. Sea $T = \{leukocyte, blood, neurone\}$ y sean:

$$\text{expand}(leukocyte) = \{leukocyte, leucocyte, white blood cell, white cell\}$$

$$\text{expand}(blood) = \{blood\}$$

$$\text{expand}(neurone) = \{neurone, neuron, nerve cell\}$$

Sea una ontología $o \in O$ con cuatro conceptos, tal que $\text{conceptnames}(o) = \{cell, white cell, neuron, tooth cell\}$. Entonces:

$$\begin{aligned} CCscore(o) &= \frac{\sum_{i=1}^n \text{matching}(t, o)}{n} \\ &= \frac{\text{matching}(leukocyte, o)}{3} + \frac{\text{matching}(blood, o)}{3} \\ &\quad + \frac{\text{matching}(neurone, o)}{3} = \frac{1+0+1}{3} = \frac{2}{3} = 0,67 \end{aligned}$$

Así, la ontología cubre los términos de entrada con un $CCscore$ de 0,67, o lo que es lo mismo, proporciona una cobertura de los términos de un 67%.

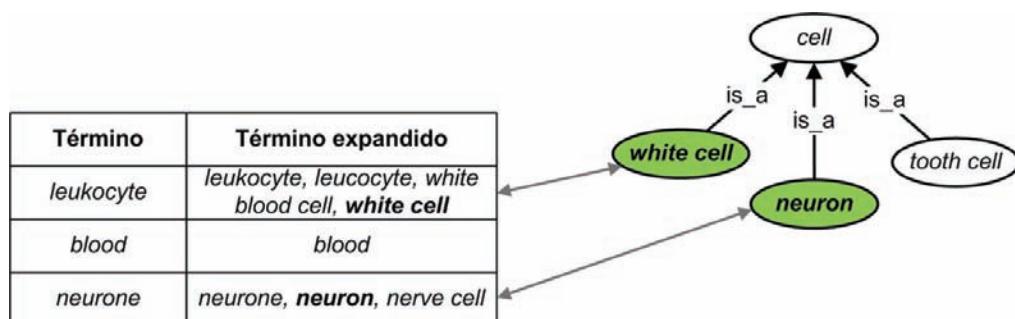


Figura 5.8. Ejemplo de cobertura proporcionada por una ontología para unos términos de entrada. Se puede ver que la ontología es capaz de cubrir 2 de los 3 términos dados.

La medida *CCscore* aquí presentada está basada en la medida conocida como *Class Match Measure* (CMM) propuesta por Alani y colegas (Alani, et al., 2006), pero adaptada para tener en cuenta únicamente correspondencias exactas. Se podría haber pensado en usar alguna medida de distancia entre cadenas de texto (e.g. distancia de Levenshtein) que permitiese tener en cuenta también correspondencias no exactas. Sin embargo, y como indican Sabou y colegas (Sabou, et al., 2006a), este mecanismo sería bastante frágil pues, aunque puede proporcionar aciertos importantes (e.g. *neurone* al buscar por *neuron*), también puede generar resultados claramente incorrectos (e.g. *update* al buscar por *date*, o *team*, *steam cell* y *teach* al buscar por *tea*). Jones y Alani (Jones & Alani, 2006) también coinciden en que considerar correspondencias parciales puede reducir considerablemente la calidad de la búsqueda.

5.4.2 Evaluación de la riqueza semántica

Las ontologías con un mayor número de elementos para un contexto determinado, se pueden considerar potencialmente más útiles para describir dicho contexto que ontologías más simples. Cuando se accede a un concepto de una ontología, resulta de utilidad encontrar cierto grado de detalle en la representación del conocimiento concerniente a tal concepto. Esto puede incluir el número de parientes del concepto, si existe una definición para él o no en la ontología, relaciones con otros conceptos (e.g. *part_of*, *located_in*, *contained_in*, etc.), etc. En base a estas características, la **evaluación de la riqueza semántica** de la ontología tratará de medir el nivel de detalle del conocimiento representado en ella.

Ejemplo 5.7. Supóngase que se desea seleccionar la mejor ontología para describir semánticamente el término *neuron*. Considerando las dos ontologías, *A* y *B*, que se muestran en la figura 5.9, se puede observar que ambas cubren dicho término. Sin embargo, la ontología *B* proporciona un conocimiento mayor del término *neuron* que la ontología *A*. La ontología *B* cuenta con una definición del término *neuron*, varios sinónimos, varios términos relacionados a través de la relación *part_of*, y varias subclases. Si el término *neuron* se describe (o anota) con la ontología *B*, esta descripción será de mayor utilidad, tanto para usuarios que deseen acceder a la ontología para consultar información sobre el concepto *neuron*, como para agentes software que traten de procesar la información de dicho concepto para realizar inferencias u otras

operaciones. Este ejemplo evidencia la importancia de considerar la riqueza semántica de la ontología durante el proceso de evaluación.

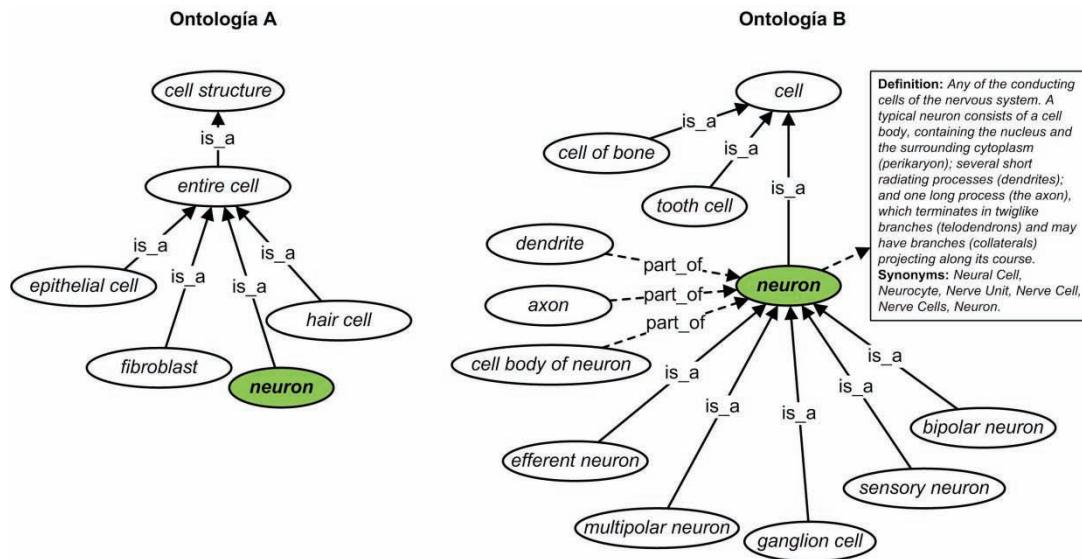


Figura 5.9. Ejemplo de dos ontologías que proporcionan distintos niveles de riqueza semántica para el concepto “neuron”.

En esta sección se presenta una medida, ideada por el autor, para calcular la riqueza semántica de una ontología, llamada *Semantic Richness Score* (*SRScore*). Dada una ontología que se ha determinado que cubre unos determinados términos de entrada, esta medida se basa en evaluar la ontología de acuerdo a los siguientes aspectos:

- Parentesco.** Para cada concepto, se calculan sus parientes directos (padres, hijos y hermanos). Este aspecto es un indicativo del nivel de detalle con el que el concepto se ha especificado.
- Información adicional.** Tiene en cuenta otra información que la ontología proporciona sobre el concepto. Se contempla la existencia de definiciones sobre el mismo, relaciones con otros conceptos, comentarios, restricciones, etc. En definitiva, aquí se valorará cualquier información sobre el concepto en la ontología, no contemplada ya en la información de parentesco.
- Conocimiento similar.** Se buscan en la ontología otros conceptos similares al concepto dado.

Dado que se desea seleccionar la ontología más adecuada a un contexto determinado, la evaluación de la riqueza semántica se debe restringir únicamente a los conceptos de la ontología implicados en la cobertura del contexto.

A continuación, se explicará en qué consiste la evaluación de los aspectos mencionados (i.e. parentesco, información adicional y conocimiento similar). Finalmente y en base a ellos, se planteará el cálculo del *SRscore*.

5.4.2.1 Evaluación de parentesco

Definición 5.11 (parientes). Sea \mathcal{O} una ontología, \mathcal{C} el conjunto de conceptos de la ontología, y $c \in \mathcal{C}$ un concepto cualquiera de \mathcal{O} . Sea P el conjunto de conceptos de \mathcal{O} que son padres directos de c , sea S el conjunto de conceptos de \mathcal{O} que son hermanos de c (teniendo en cuenta todos los padres de c) y sea H el conjunto de conceptos de \mathcal{O} que son hijos directos de c . Se define la función $relatives: \mathcal{C} \rightarrow \mathbb{N}$ como sigue:

$$relatives(c) = |P| + |S| + |H|$$

Es decir, para un concepto, la función *relatives* cuenta el número total de padres, hijos y hermanos directos del concepto.

Ejemplo 5.8. Considérense de nuevo las ontologías de la figura 5.9. Sea c_1 el concepto *neuron* de la ontología \mathcal{A} , y c_2 el concepto *neuron* de la ontología \mathcal{B} . Entonces:

$$relatives(c_1) = 1 + 3 + 0 = 4$$

$$relatives(c_2) = 1 + 2 + 5 = 8$$

A partir de la función *relatives* se calcula el **índice de parentesco** (*relatives index*) de una ontología respecto a un conjunto de términos representados por ella.

Definición 5.12 (índice de parentesco). Sea \mathcal{O} una ontología y \mathcal{C} el conjunto de todos los conceptos de \mathcal{O} . Se define el índice de parentesco (*relatives index*) para cualquier subconjunto de conceptos de la ontología $D \subseteq \mathcal{C}$ como una función $ri: \mathcal{P}(\mathcal{C}) \rightarrow [0, +\infty)$, teniendo en cuenta que $\mathcal{P}(\mathcal{C})$ es el conjunto de todos los posibles subconjuntos de conceptos de la ontología \mathcal{O} . Entonces:

$$ri(D) = \frac{\sum_{i=1}^n relatives(d_i)}{n} \quad \forall d_i \in D : 1 \leq i \leq n \wedge n = |D|$$

El índice de parentesco indica el promedio de parientes directos de los conceptos de una ontología que cubren los términos de entrada.

Ejemplo 5.9. Considérese que o es la ontología B de la figura 5.9, y que $D = \{d_1, d_2, d_3\}$ es un subconjunto de los conceptos de B , tal que:

$$\text{conceptname}(d_1) = \text{cell}$$

$$\text{conceptname}(d_2) = \text{axon}$$

$$\text{conceptname}(d_3) = \text{neuron}$$

En este caso, el índice de parentesco de D se calcularía de la siguiente manera:

$$\begin{aligned} ri(D) &= \frac{\text{relatives}(d_1)}{3} + \frac{\text{relatives}(d_2)}{3} + \frac{\text{relatives}(d_3)}{3} \\ &= \frac{3 + 0 + 8}{3} = \frac{11}{3} = 3,67 \end{aligned}$$

5.4.2.2 Evaluación de información adicional

Definición 5.13 (información adicional). Sea C el conjunto de conceptos de una ontología o , la función $\text{additionalinf}: C \rightarrow \mathbb{N}$ calcula el número características de un concepto c que proporcionan información sobre dicho concepto, exceptuando:

- La información proporcionada por las relaciones jerárquicas (is_a), ya considerada en la evaluación de parentesco.
- La información sobre la etiqueta (nombre) del concepto. Se asume que todo concepto tendrá siempre un nombre, y no se considerará ésta una característica diferenciadora.
- La información sobre las posibles instancias del concepto. Se ha decidido no tener en cuenta esta información para evitar valorar en exceso ontologías superpobladas (con muchas instancias), considerando además que, en general, la información más valiosa de una ontología la constituye la estructura de su modelo conceptual (conceptos y relaciones entre ellos) y no sus instancias.

Así, en esta información adicional se tendrían en cuenta relaciones del concepto con otros conceptos, definiciones, sinónimos del nombre del concepto, restricciones sobre posibles valores o tipos de datos que puede adoptar el concepto, etc.

Ejemplo 5.10. La clase *neuron* de la ontología *B* de la figura 5.9, posee 1 definición, 6 sinónimos y 3 clases vinculadas a través de la relación *part_of*. En este caso: $additionalinf(neuron) = 1 + 6 + 3 = 10$.

Tras cuantificar la información adicional que proporciona una clase determinada, se propone el cálculo de un índice que revela la información adicional para un subconjunto de los conceptos de una ontología.

Definición 5.14 (índice de información adicional). Sea o una ontología y C el conjunto de todos los conceptos de o . Se define el índice de información adicional (*additional information index*) para cualquier $D \subseteq C$ como una función $ai: \mathcal{P}(C) \rightarrow [0, +\infty)$ tal que:

$$ai(D) = \frac{\sum_{i=1}^n additionalinf(d_i)}{n} \quad \forall d_i \in D : 1 \leq i \leq n \wedge n = |D|$$

5.4.2.3 Evaluación de conocimiento similar

En dominios específicos como la biomedicina, es habitual encontrar ontologías que, además de contener un concepto determinado, proporcionan otros muchos conceptos similares a él, que no tienen por qué encontrarse directamente relacionados. Cuando se evalúa la riqueza que proporciona una ontología para un contexto, también es importante tener en cuenta estos conceptos relacionados o similares.

Ejemplo 5.11. Supóngase que se desea seleccionar la mejor ontología para describir semánticamente el término *heart*. Considerando las ontologías *A* y *B* de la figura 5.10 se puede observar que ambas cubren el término, pues ambas contienen el concepto *heart*. Sin embargo, la ontología *B* proporciona un mayor conocimiento acerca del contexto del término *heart*, ya que contiene otros 6 conceptos cuyo significado pertenece al dominio de la cardiología, mientras que la ontología *A* no contiene ninguno. Debido a esto, se puede decir que la ontología *B* cuenta con más potencial que *A* para describir dicho ámbito.

La fase de evaluación de conocimiento similar trata de medir cuánto conocimiento posee la ontología, en el ámbito del contexto en el que está siendo evaluada. Para esto, se cuentan el número de conceptos de la ontología cuyo nombre contiene al nombre del concepto que se está evaluando, o al nombre de alguno de los sinónimos de este

concepto. En definitiva, se trata de buscar las correspondencias de subcadena (*substring matchings*) en la ontología, para el nombre del concepto dado y todos sus sinónimos. A continuación, se explica esta noción de manera más formal.

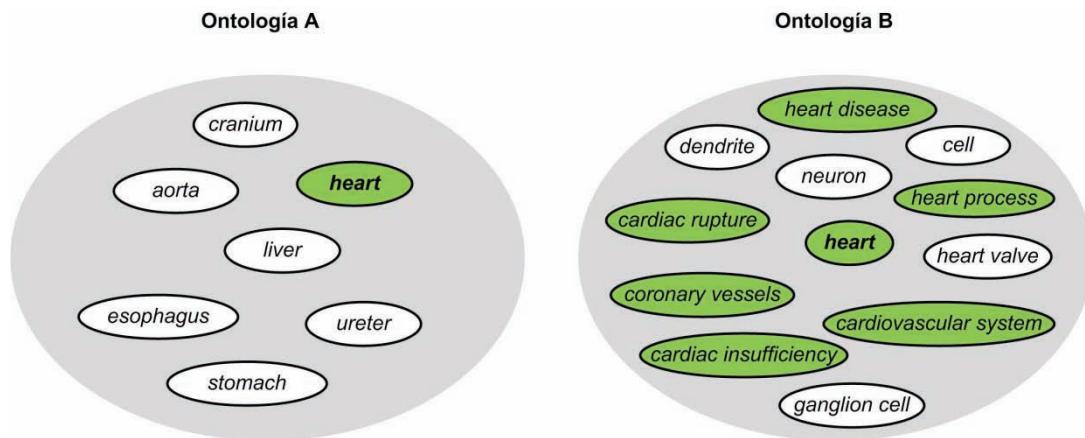


Figura 5.10. Ejemplo de conjuntos de conceptos de dos ontologías *A* y *B*. Se ha omitido la representación de las relaciones entre entidades.

Definición 5.15 (subcadenas). Sea \mathbb{S} el conjunto de todas las posibles cadenas de texto sobre un alfabeto. Sea $t \in \mathbb{S}$ una cadena de texto, se define la función *substrings*: $\mathbb{S} \rightarrow \mathcal{P}(\mathbb{S})$ como aquella función que, para una cadena de texto determinada, devuelve el conjunto formado por todas sus posibles subcadenas, exceptuando la cadena vacía. Se dirá que una cadena w es una subcadena de otra cadena z , si existen las cadenas x e y para las cuales $z = xwy$.

Ejemplo 5.12. Para cadena *cell*:

$$\text{substrings}(cell) = \{cell, cel, ce, c, e, l, el, ell, ll\}$$

En base a esta función, se propone calcular los conceptos de una ontología similares a un concepto dado, como se explica a continuación.

Definición 5.16 (conocimiento similar). Sea C el conjunto de todos los conceptos de una ontología, sea $c \in C$ uno de estos conceptos. Se define la función *similarknowledge*(c): $C \rightarrow \mathbb{N}$ como:

$$\text{similarknowledge}(c) =$$

$$\begin{aligned} & |\{c_i \in C \mid c_i \neq c \wedge \exists s \in \mathbb{S} \mid s \in \text{expand}(\text{conceptname}(c)) \wedge s \\ & \quad \in \text{substrings}(\text{conceptname}(c_i))\}| \end{aligned}$$

Ejemplo 5.13. Sean a y b dos conceptos de las ontologías \mathcal{A} y \mathcal{B} de la figura 5.10, respectivamente, tal que $\text{conceptname}(a) = \text{conceptname}(b) = \text{heart}$. Considérese también que el conjunto de términos obtenidos tras realizar la expansión semántica del término heart es el siguiente:

$$\text{expand}(\text{heart}) = \{\text{heart}, \text{hearts}, \text{cardio}, \text{cardiac}, \text{coronary}\}$$

Entonces:

$$\text{similarknowledge}(a) = |\emptyset| = 0$$

$$\text{similarknowledge}(b) = |\{b_1, b_2, b_3, b_4, b_5, b_6\}| = 6, \text{ con:}$$

$$\text{conceptname}(b_1) = \text{heart disease}$$

$$\text{conceptname}(b_2) = \text{cardiac rupture}$$

$$\text{conceptname}(b_3) = \text{heart process}$$

$$\text{conceptname}(b_4) = \text{coronary vessels}$$

$$\text{conceptname}(b_5) = \text{cardiovascular system}$$

$$\text{conceptname}(b_6) = \text{cardiac insufficiency}$$

Definición 5.17 (índice de conocimiento similar). Sea \mathcal{O} una ontología y \mathcal{C} el conjunto de todos los conceptos de \mathcal{O} . Se define el índice de conocimiento similar (*similar knowledge index*) para cualquier $D \subseteq \mathcal{C}$ como una función $si: \mathcal{P}(\mathcal{C}) \rightarrow [0, +\infty)$ tal que:

$$si(D) = \frac{\sum_{i=1}^n \text{similarknowledge}(d_i)}{n} \quad \forall d_i \in D : 1 \leq i \leq n \wedge n = |D|$$

5.4.2.4 Cálculo del SRscore

Tras realizar la evaluación del nivel de parentesco, información adicional y conocimiento similar que proporciona una ontología para un contexto determinado, estas tres medidas se combinan para dar lugar a un único valor que representa la riqueza semántica de la ontología en el contexto representado mediante los términos de entrada. Este valor se ha denominado *SRscore*, y se define a continuación.

Definición 5.18 (SRscore). Sea T un conjunto de términos resultantes tras el proceso de normalización y corrección. Sea \mathcal{O} una ontología y \mathcal{C} el conjunto de todos

los conceptos de σ . Sea $D \subseteq C$ un subconjunto de los conceptos de la ontología σ , cada uno de los cuales cubre un término $t \in T$. Se define la puntuación de riqueza semántica o *SRscore* (*Semantic Richness Score*) como una función $SRscore : \mathcal{P}(C) \rightarrow [0, 1]$, tal que:

$$SRscore(D) = w_r * norm(ri(D)) + w_a * norm(ai(D)) + w_s * norm(si(D))$$

Como se ha visto, el rango de las funciones $ri(D)$, $ai(D)$ y $si(D)$ es el intervalo $[0, +\infty)$. Debido a esto, y para obtener un resultado final del *SRscore* en el intervalo $[0, 1]$, se ha realizado una normalización (función *norm*) basada en la distribución de los valores a partir de los que se calcula cada índice, para un conjunto de ontologías del dominio. En el apartado 5.4.5 se explica en qué ha consistido esta normalización. Además, en el apartado 5.4.6 se explicará la utilidad de los pesos w_r , w_a y w_s , y se propondrá una forma de calcularlos.

5.4.3 Evaluación de la popularidad

A parte de evaluar en qué medida una ontología cubre un contexto determinado, y la riqueza semántica de la ontología en relación a ese contexto, existe otro aspecto que requiere una atención especial. ¿Qué ocurre si la información de una ontología es errónea? De acuerdo a la definición de ontología proporcionada por Studer (Studer, et al., 1998), que se puede considerar una de las definiciones de ontología más precisas y completas hasta el momento, “*una ontología captura conocimiento consensuado, es decir, no es un conocimiento privado de algún individuo, sino aceptado por un grupo*”. Teniendo esto en cuenta, un método ideado para evaluar una ontología debería tener en cuenta el nivel de aceptación colectiva, o popularidad de la ontología.

En este trabajo se propone, como idea novedosa, un método para la **evaluación de la popularidad** de una ontología, basado en el conocimiento colectivo almacenado en recursos Web 2.0, cuyo valor se debe a la agregación de múltiples contribuciones individuales. Esta evaluación consiste en calcular una medida, llamada *Popularity Score* (*Pscore*), que consiste en contar el número de referencias a la ontología desde un conjunto de recursos Web 2.0 predefinidos. Dependiendo del tipo de recurso, el criterio para medir las referencias puede ser diferente. A continuación, se define esta medida.

Definición 5.19 (*Pscore*). Sea o una ontología, sea $R = \{r_1, r_2, \dots, r_n\}$ el conjunto de recursos Web 2.0 tenidos en cuenta para evaluar la popularidad de o , sea $W = \{w_1, w_2, \dots, w_n\}$ el conjunto de pesos en el intervalo $[0, 1]$ para los recursos de R , sea $refs : O \times R \rightarrow \mathbb{N}$ la función que cuenta el número de referencias a una ontología o desde un recurso $r \in R$, y sea $norm : [0, +\infty) \rightarrow [0, 1]$ una función de normalización (definida en el apartado 5.4.5), entonces:

$$Pscore(o, R) = \sum_{i=1}^n w_i \cdot norm(refs(o, r_i)) \quad \forall r_i \in R : 1 \leq i \leq n \wedge n = |R|$$

En el caso de la aproximación propuesta, se ha decidido usar los siguientes recursos:

1. **BioPortal**⁴⁹. Es el recurso de referencia para la publicación de ontologías biomédicas. Actualmente contiene más de 200 ontologías del ámbito de la biomedicina. En este caso, la función $refs$ proporcionará un valor de 1 si la ontología se encuentra indexada en bioportal, o un 0 en caso contrario.
2. **PubMed**⁵⁰. Es un recurso Web que permite acceder a millones de publicaciones científicas de los ámbitos médico y biomédico. Ontologías que llegan a ser referenciadas en artículos científicos son ontologías generalmente maduras, que han pasado un proceso de evaluación técnica y, en definitiva, de fiar. Para este recurso, se cuentan el número de artículos en PubMed en los que aparece, al menos una vez, el nombre de la ontología.
3. **Wikipedia**⁵¹. Se trata de un esfuerzo colectivo por crear una enciclopedia gratuita, libre y accesible para todo el mundo. De forma similar a como se hace para el recurso PubMed, en este caso se contabilizan el número de artículos en Wikipedia en los que figura, como mínimo una vez, el nombre de la ontología.
4. **Twitter**⁵². Es un sitio Web que permite a sus usuarios enviar y leer entradas de texto de una longitud máxima de 140 caracteres. Cada una de estas entradas se conoce como *tweet*. Para calcular el número de referencias a una ontología en

⁴⁹ <http://bioportal.bioontology.org/>

⁵⁰ <http://www.ncbi.nlm.nih.gov/pubmed>

⁵¹ <http://www.wikipedia.org/>

⁵² <http://twitter.com/>

Twitter, se propone contar el número de *tweets* publicados que contienen, al menos una vez, el nombre de la ontología.

Teniendo en cuenta estos recursos, el cálculo del *Pscore* para una ontología se realizaría de acuerdo a la siguiente expresión:

$$\begin{aligned} Pscore(o, R) = & w_{bioportal} \text{ norm} \left(\text{refs}(o, r_{bioportal}) \right) \\ & + w_{pubmed} \text{ norm} \left(\text{refs}(o, r_{pubmed}) \right) \\ & + w_{wikipedia} \text{ norm} \left(\text{refs}(o, r_{wikipedia}) \right) \\ & + w_{twitter} \text{ norm} \left(\text{refs}(o, r_{twitter}) \right) \end{aligned}$$

Donde:

- $\text{refs}(o, r_{bioportal})$ es el número de referencias a la ontología o desde BioPortal.
- $\text{refs}(o, r_{pubmed})$ es el número de referencias a la ontología o desde PubMed.
- $\text{refs}(o, r_{wikipedia})$ es el número de referencias a la ontología o desde Wikipedia.
- $\text{refs}(o, r_{twitter})$ es el número de referencias a la ontología o desde Twitter.

Al igual que para el cálculo del *SRscore*, la fórmula propuesta hace uso de la función *norm*, que se explicará en detalle en el apartado 5.4.5 y de un conjunto de pesos, cuyo ajuste se explicará en el apartado 5.4.6.

5.4.4 Agregación de puntuaciones

Finalmente, y una vez se dispone de las puntuaciones de cobertura del contexto, riqueza semántica y popularidad de una ontología, estos valores se agregan en una única medida, denominada **puntuación agregada** (*aggregated score*), cuyo rango es [0,1], y que se calcula como se explica a continuación.

Definición 5.20 (puntuación agregada). Sea T el conjunto de términos de entrada resultantes tras el proceso de normalización y corrección. Sea o una ontología, y C el conjunto de todos sus conceptos. Sea $D \subseteq \mathcal{P}(C)$ un subconjunto de los conceptos de la ontología o , cada uno de los cuales cubre un término $t \in T$. Se define la puntuación agregada, o puntuación final para la ontología o en el contexto representado por los términos de T como:

$$\text{aggregatescore}(o, T) = w_{CC} * CCscore(o, T) + w_{SR} * SRscore(D) + w_P * Pscore(o)$$

La agregación de las medidas de evaluación se realiza utilizando tres pesos: w_{CC} , w_{SR} y w_P , que permiten establecer la influencia de cada medida de evaluación (i.e. cobertura del contexto, riqueza semántica y popularidad, respectivamente) en la puntuación final. El ajuste de estos pesos se tratará en el apartado 5.4.6.

5.4.5 Normalización de valores

Como se ha visto, la aproximación depende de una función $norm$, cuya finalidad es trasladar un valor en el intervalo abierto $[0, +\infty)$ al intervalo $[0, 1]$. Esta función se define de la forma siguiente:

Definición 5.21 (normalización). Sea v una variable, que toma valores en el intervalo $[0, +\infty)$, se define la función de normalización $norm : [0, +\infty) \rightarrow [0, 1]$ como aquella función que, dado un valor cualquiera de v , devuelve el valor correspondiente en el intervalo $[0, 1]$. Este valor se obtiene tras aplicar una discretización a intervalos de igual frecuencia (Liu et al., 2002) a los valores de v , y asignar a cada clase obtenida un valor discreto en el intervalo $[0, 1]$. El cálculo de esta función se realiza siguiendo los siguientes pasos:

1. Selección de una muestra de referencia (e.g. número de referencias desde un recurso Web a las ontologías del repositorio).
2. Eliminación del 0 en la muestra. Para el 0, la salida será siempre 0.
3. División de la muestra en intervalos de igual frecuencia:
 - a. Número de intervalos: una decisión importante es determinar el número de intervalos en el que se dividirá la muestra. Existen varias aproximaciones para determinar este número de intervalos. En este caso se ha decidido utilizar la *regla de Sturges* (Daniel & Wayne, 2009), según la que el número de intervalos se calcula de la siguiente manera: $nint = 1 + 3,22 * \log(n)$, siendo n el tamaño muestral. En todo caso, se ha adoptado un $nint$ máximo de 20.
 - b. Cálculo de los puntos límite de los intervalos: si se da el caso de que existen muchas ocurrencias de un determinado valor, es posible que se

obtengan varios intervalos coincidentes. En estos casos, los intervalos solapados se agrupan en un único intervalo.

4. Asignación a cada intervalo en el rango $[0, +\infty)$ de un valor en el intervalo $[0, 1]$. Al i -ésimo intervalo se le asignará el valor i/k , con $1 \leq i \leq k$, donde k es el número de intervalos en que se ha dividido la muestra.

Ejemplo 5.14. Supóngase que las referencias en el recurso PubMed para un conjunto ficticio de 10 ontologías son las que se muestran en la tabla 5.11.

Tabla 5.11. Referencias desde un recurso Web a diez ontologías ficticias.

Ontología	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}
Nº Referencias	4	21	2	0	3215	0	3	1	6	47

Una normalización lineal de la muestra (e.g. dividir por el valor máximo), provocaría la pérdida de la información diferenciadora que aportan los valores más bajos. La función *norm* permitirá obtener, para cada valor de la muestra, un valor en el intervalo $[0, 1]$, conservando la capacidad de discriminación de los valores originales. Para esto, se realizan los siguientes pasos:

1. Eliminación de ceros. Si una ontología tiene 0 referencias, se le asignará un valor de popularidad para el recurso dado, igual a 0. La muestra resultante, ordenada de forma creciente, es: $\{1, 2, 3, 4, 6, 21, 47, 3215\}$.
2. División de la muestra en intervalos de igual frecuencia. El nº de intervalos más adecuado según la *regla de Sturges* se calcula de la siguiente manera:

$$nint = 1 + 3,22 * \log(n) = 1 + 3,22 * \log(8) = 3,90 \cong 4$$

La muestra se dividirá en 4 intervalos, cada uno de los cuales contendrá el mismo número de ocurrencias (intervalos con la misma frecuencia). En este caso, los puntos límite de los intervalos vendrían dados por la posición de los percentiles 25, 50 y 75 (i.e. cuartiles). Estos puntos serían: $\{2,5; 5; 34\}$. Estos puntos dividen la muestra en cuatro intervalos de clase, que son los siguientes:

$$int_1: \text{para } 0 < x < 2,5$$

$$int_2: \text{para } 2,5 \leq x < 5$$

int_3 : para $5 \leq x < 34$

int_4 : para $34 \leq x$

3. A cada intervalo de clase, se le asigna un valor en el intervalo $[0, 1]$, que se calcula como $\frac{i}{k}$, con $1 \leq i \leq k$, donde k es el número de intervalos en que se ha dividido la muestra. Es decir, para este caso, la función *norm* para un valor $x \in [0, +\infty)$ se calcularía de la siguiente manera:

$$norm(x) = \begin{cases} 0; & \text{si } x = 0 \\ 1/4 = 0,25; & \text{si } 0 < x < 2,5 \\ 2/4 = 0,5; & \text{si } 2,5 \leq x < 5 \\ 3/4 = 0,75; & \text{si } 5 \leq x < 34 \\ 4/4 = 1; & \text{si } 34 \leq x \end{cases}$$

De esta manera, se podría convertir cualquier valor en el rango $[0, +\infty)$ al rango $[0, 1]$. Por ejemplo:

$$\text{Si } refs(o, r_{pubmed}) = 15 \text{ entonces } norm(refs(o, r_{pubmed})) = 0,75$$

$$\text{Si } refs(o, r_{pubmed}) = 675 \text{ entonces } norm(refs(o, r_{pubmed})) = 1$$

5.4.6 Ajuste de pesos

Como se ha visto en los anteriores apartados, la aproximación contempla el uso de varios pesos, que permiten dar mayor o menor importancia a los diferentes criterios utilizados para evaluar cada una de las ontologías candidatas. Estos pesos son los siguientes:

- Pesos utilizados para calcular la riqueza semántica de cada ontología (ver apartado 5.4.2):
 - w_r : Peso del índice de parentesco de un concepto.
 - w_a : Peso del índice de información adicional de un concepto.
 - w_s : Peso del índice de conocimiento similar de un concepto.
- Pesos utilizados para calcular la popularidad de una ontología (ver apartado 5.4.3):
 - $w_{bioportal}$: Peso del recurso Web BioPortal.
 - w_{pubmed} : Peso del recurso Web PubMed.
 - $w_{wikipedia}$: Peso del recurso Web Wikipedia.

- w_{twitter} : Peso del recurso Web Twitter.
- Pesos utilizados para realizar la agregación de las puntuaciones obtenidas por cada criterio de evaluación (ver apartado 5.4.4):
 - w_{CC} : Peso de la evaluación de la cobertura del contexto.
 - w_{SR} : Peso de la evaluación de la riqueza semántica.
 - w_P : Peso de la evaluación de la popularidad.

Estos pesos pueden tomar un valor en el intervalo [0, 1], y están sujetos a las siguientes restricciones:

$$w_r + w_a + w_s = 1$$

$$w_{\text{bioportal}} + w_{\text{pubmed}} + w_{\text{wikipedia}} + w_{\text{twitter}} = 1$$

$$w_{CC} + w_{SR} + w_P = 1$$

El ajuste de estos pesos, afectará directamente a los resultados obtenidos por la aproximación, y tendrá que ser realizado teniendo en cuenta las características del entorno en el que la aproximación se vaya a implantar, y del uso final que vayan a recibir las ontologías. Por ejemplo, en caso de contar con un repositorio de ontologías fiable, formado por ontologías consensuadas por expertos, se podría dar menor peso al criterio de popularidad (i.e. reducir el valor de w_P). Sin embargo, trabajando con un repositorio en el que cualquier usuario puede incorporar cualquier ontología sin ningún proceso de revisión, sería recomendable dar mayor importancia al criterio de popularidad para tratar de evitar la sobrevaloración de ontologías no relevantes para la comunidad biomédica.

Para evaluar el prototipo de sistema de selección que se ha construido en esta tesis, el ajuste de estos pesos se ha realizado en base a la opinión de 5 expertos en el ámbito de las ontologías biomédicas. Para ello, se ha diseñado un *Formulario de ajuste de pesos*, que se puede consultar en el anexo II de este documento. Los detalles de este proceso y los resultados obtenidos se presentarán en el apartado 6.4.6.

5.5 Combinación y ordenación de ontologías

Como explican Sabou y colegas (Sabou, et al., 2006b), la salida del proceso de selección se puede expresar de diversas maneras. Las ontologías seleccionadas pueden presentarse como una lista ordenada de ontologías. También, la selección puede devolver posibles combinaciones de ontologías que en conjunto satisfacen una determinada necesidad de información. En general, es importante proporcionar una salida “amigable” para el usuario.

La aproximación presentada en esta tesis proporciona dos tipos de salida:

- **Salida simple.** Esta salida consiste en un ranking, en el que cada elemento es una ontología. El ranking se realiza en base a la puntuación final obtenida durante el proceso de evaluación.
- **Salida combinada.** En esta salida, cada elemento del ranking puede estar formado por una o por varias ontologías. El principal propósito de esta salida es incrementar la cobertura del contexto, a través de la combinación de varias ontologías que cubren términos diferentes. Es decir, si una ontología A cubre 3 términos de entrada, y una ontología B cubre 5 términos de entrada, diferentes a los términos que cubre A , ambas ontologías juntas proporcionan una cobertura de 8 términos. Este tipo de salida resulta de gran utilidad, especialmente en entornos en que la cobertura del contexto es crítica (e.g. anotación semántica de campos de una BD, para su futura integración). La salida combinada se calcula a partir de las puntuaciones obtenidas tras evaluar cada ontología de forma individual, realizando los siguientes pasos:
 1. **Obtención de todas las posibles combinaciones** sin repetición de las ontologías.
 2. **Cálculo de puntuaciones combinadas.** Para cada combinación de ontologías se calcula:
 - Cobertura del contexto ($CCscore$) combinada: se calcula en base a los términos que cubren las ontologías de forma conjunta. Si existen términos cubiertos por varias ontologías, se elige aquella de mayor riqueza semántica.

- Riqueza semántica (*SRscore*) y popularidad (*Pscore*) combinadas: se calculan en base a los valores de riqueza semántica y popularidad de cada ontología por separado, proporcionalmente al número de términos que cubre cada ontología.
 - Puntuación final. Se calcula de la misma forma que para una ontología individual, pero a partir del *CCscore*, *SRscore* y *Pscore* combinados.
3. **Ordenación.** Una vez se cuenta con la puntuación final para todas las posibles combinaciones de las ontologías, la ordenación se realiza de acuerdo a estas puntuaciones. En caso de empate, tienen prioridad las combinaciones de ontologías que mayor cobertura del contexto proporcionan. En caso de un nuevo empate, tendrían prioridad las combinaciones en las que intervienen un menor número de ontologías. Se pretende obtener las mejores puntuaciones minimizando en la medida de lo posible el número de ontologías de cada resultado.

A continuación se presenta un sencillo ejemplo que permite entender en qué consiste la salida simple y combinada.

Ejemplo 5.15. Supóngase que se desea seleccionar las mejores ontologías para siguiente conjunto de entrada:

$$T = \{white\ cell, chemotherapy, melanoma, stomach, cavity\ of\ stomach\}$$

Y supóngase también que el repositorio de ontologías contiene únicamente dos ontologías: el NCI Thesaurus, que cubre todos los términos de *T* excepto *cavity of stomach*, y la ontología Foundational Model of Anatomy, que cubre todos los términos excepto *chemotherapy* y *melanoma*. Para este caso, la salida simple se muestra en la tabla 5.12 , mientras que la salida combinada se puede ver en la tabla 5.13.

Tabla 5.12. Salida simple para los datos del ejemplo.

Posición	Ontología	CCscore	SRscore	Pscore	Score final
1	NCI Thesaurus	0,800	0,934	0,863	0,847
2	Foundational Model of Anatomy	0,600	0,885	0,453	0,646

Tabla 5.13. Salida combinada para los datos del ejemplo.

Posición	Ontologías	CCscore	SRscore	Pscore	Score final
1	NCI Thesaurus Foundational Model of Anatomy	1,000	0,924	0,781	0,939
2	NCI Thesaurus	0,800	0,934	0,863	0,847
3	Foundational Model of Anatomy	0,600	0,885	0,453	0,646

Se puede ver que en la salida simple, la ontología NCI Thesaurus sería la primera en el ranking, con una cobertura del contexto de un 80% y una puntuación final de 0,847. En la salida combinada, la combinación de las dos ontologías permite alcanzar una cobertura total del contexto (100%). En este caso, en NCI Thesaurus cubriría 4 de los 5 términos de entrada, mientras que el Foundational Model of Anatomy cubriría únicamente uno (*cavity of stomach*). A pesar de que esta última ontología podría cubrir también los términos *white cell* y *stomach*, se decide que éstos sean cubiertos por el NCI Thesaurus, por tener mayor riqueza semántica (0,934 respecto a 0,885). De esta manera, en la salida combinada el NCI Thesaurus cubre un 80% de los términos de entrada y el Foundational Model of Anatomy un 20%. Estos porcentajes se usan para calcular los valores combinados de *SRscore* y *Pscore*. El nuevo *SRscore* será un 80% del obtenido por el NCI Thesaurus individualmente, más un 20% del obtenido por Foundational Model of Anatomy, es decir:

$$SRscore = \frac{80 \cdot 0,934}{100} + \frac{20 \cdot 0,885}{100} = 0,924$$

Y el nuevo *Pscore* se calcularía de la misma forma pero usando los valores correspondientes a la popularidad:

$$Pscore = \frac{80 \cdot 0,863}{100} + \frac{20 \cdot 0,453}{100} = 0,781$$

5.6 Diferencias con las aproximaciones existentes

La aproximación para la selección de ontologías que se propone se ha comparado con las aproximaciones existentes, utilizando los criterios presentados en el apartado 3.4.3. Esta comparativa ha permitido identificar varias características de la aproximación, que

la diferencian del trabajo realizado hasta el momento. Estas características diferenciadoras se pueden intuir fácilmente observando la tabla 5.14. A continuación, se resaltan las más relevantes:

- Es la única aproximación propuesta hasta el momento que contempla proporcionar combinaciones de ontologías como resultado. Ésta se puede considerar una ventaja especialmente relevante respecto al trabajo existente si se tiene en cuenta que varios autores la han considerado como una característica esencial de los sistemas de selección de ontologías (Sabou, et al., 2006b).
- Es una de las cuatro aproximaciones (una de dos en el dominio biomédico) que contempla evaluar la popularidad de las ontologías candidatas, y la única capaz de evaluar dicha popularidad sin solicitar a los usuarios valoraciones explícitas sobre las ontologías. Como ya se ha explicado, la aproximación plantea evaluar la popularidad en base al conocimiento existente en varios recursos Web creados a partir del conocimiento de múltiples usuarios.
- Actualmente, sólo existen tres aproximaciones completamente automáticas para la selección de ontologías biomédicas, y la aproximación propuesta en este trabajo es una de ellas.
- Considerando el número de criterios de selección contemplados, se puede considerar la aproximación más completa hasta el momento, ya que tiene en cuenta 8 de los 12 criterios propuestos. En el ámbito biomédico le sigue el Recomendador del NCBO (con 5 de 12).
- Es la única de las aproximaciones de selección de ontologías biomédicas que realiza desambiguación de los términos de entrada, y una de las dos aproximaciones en el dominio biomédico que realiza expansión semántica de las consultas.

Además de estas características diferenciadoras, también resulta de interés destacar los siguientes aspectos:

- A nivel teórico, la aproximación se podría plantear como una aproximación genérica, aplicable a cualquier dominio. Sin embargo, la aplicación a otros dominios dependerá principalmente de la existencia de un recurso

terminológico específico del dominio que permita obtener buenos resultados durante la fase de expansión semántica (i.e. un recurso equivalente a UMLS en otros dominios).

- La aproximación presentada es independiente del lenguaje en el que la ontología se encuentra representada (e.g. OWL, RDF, OBO, etc.), pues se basa en nociones comunes a todas las ontologías y no específicas de cada lenguaje de representación.
- También es importante considerar que la aproximación es dependiente en gran medida de los ajustes de pesos realizados, y éstos ajustes dependerán del entorno en el que la aproximación se vaya a implantar y del uso final que vayan a recibir las ontologías. Por ejemplo, en casos en que se cuente con un repositorio de ontologías muy fiable, se podrá dar menor peso al criterio de popularidad, y en casos en los que sea imprescindible obtener una elevada cobertura del contexto, sería recomendable incrementar el peso correspondiente a este criterio.

Tabla 5.14. Comparativa de la aproximación para la selección de ontologías propuesta (destacada en color azul) con las aproximaciones existentes. Las celdas en color verde indican un valor positivo para el criterio correspondiente, mientras que las celdas en naranja indican un valor negativo. Las celdas blancas indican indeterminado o no aplicable. Se resaltan las aproximaciones de selección del ámbito biomédico.

		Criterio												
		Aproximación												
Dominio biomédico	Búsqueda	Swoogle (Ding, et al., 2004)	Automatización	Dinamismo	Corresp. de términos	Corresp. propiedades	Expans. de consultas	Medidas estructurales	Conectividad	Desambiguación	Razonamiento	Popularidad	Metadatos	Salida combinada
		Watson (d'Aquin, et al., 2007)												
Dominio biomédico	Selección	OntoSearch (Zhang, et al., 2004)												
		OntoSearch2 (Pan, et al., 2006)												
Dominio biomédico	Selección	OntoKhoj (Patel, et al., 2003)												
		Búsqueda de BioPortal (Whetzel, et al., 2009)												
Dominio biomédico	Selección	EBI Ontology Lookup Service (Côté, et al., 2006)												
		Usuarios de BioPortal (Noy, et al., 2009)												
Dominio biomédico	Selección	Tan & Lambrix (Tan & Lambrix, 2009)												
		Maiga (Maiga, 2009)												
Dominio biomédico	Selección	Alani et al. (Alani, et al., 2007)												
		Recomendador del NCBO (Jonquet, et al., 2010)												
Dominio biomédico	Selección	Aproximación propuesta												
		Brewster et al. (Brewster, et al., 2004)												
Dominio biomédico	Selección	(ONTO) ² Agent (Arpírez, et al., 2000)												
		DL-AOSF (Wang, et al., 2008)												
Dominio biomédico	Selección	OntoSelect (Buitelaar, et al., 2004)												
		ONTOMETRIC (Lozano-Tello & Gómez-Pérez, 2004)												
Dominio biomédico	Selección	AKTiveRank (Alani, et al., 2006)												
		OntoQA (Tartir & Arpinar, 2007)												
Dominio biomédico	Selección	CombiSQORE (Ungrangsi, et al., 2008)												
		Hong et al. (Hong, et al., 2005)												
Dominio biomédico	Selección	WebCORE (Cantador, et al., 2007)												
		Sabou et al. (Sabou, et al., 2006a)												
Dominio biomédico	Selección	Supekar, 2005 (Supekar, 2005)												
		Lewen et al. (Lewen, et al., 2006)												

6 Implementación de la aproximación propuesta

La aproximación que se ha presentado en el capítulo anterior no impone restricciones acerca de la tecnología utilizada para construir un sistema de selección de ontologías biomédicas. Existen diferentes alternativas tecnológicas que se pueden utilizar para construir una instancia de dicha aproximación. En este capítulo, se describe una posible implementación, que ha dado lugar al prototipo para la selección de ontologías biomédicas denominado BiOSS (Biomedical Ontology Selection System)⁵³.

6.1 Descripción general

Como explican Cantador y colegas (Cantador, et al., 2007), existen dos escenarios fundamentales para la reutilización de ontologías. El primero se refiere al caso en el que un usuario afronta el problema de encontrar la ontología u ontologías más adecuadas para describir un determinado dominio. El segundo escenario tiene que ver con la situación en que las aplicaciones de Web Semántica necesitan encontrar de forma automática y dinámica una ontología.

Para poder satisfacer los requerimientos de ambos escenarios, y como se puede ver en la figura 6.1, el sistema BiOSS dispone de una interfaz de Servicio Web, que hace posible que éste pueda ser ejecutado desde entornos de reutilización automática de ontologías (e.g. por parte de un agente software). Adicionalmente, sobre la interfaz de

⁵³ Disponible en <http://bioss.ontologyselection.com/>

Servicio Web se ha desarrollado una interfaz gráfica Web, que un usuario puede utilizar para llevar a cabo el proceso de selección.

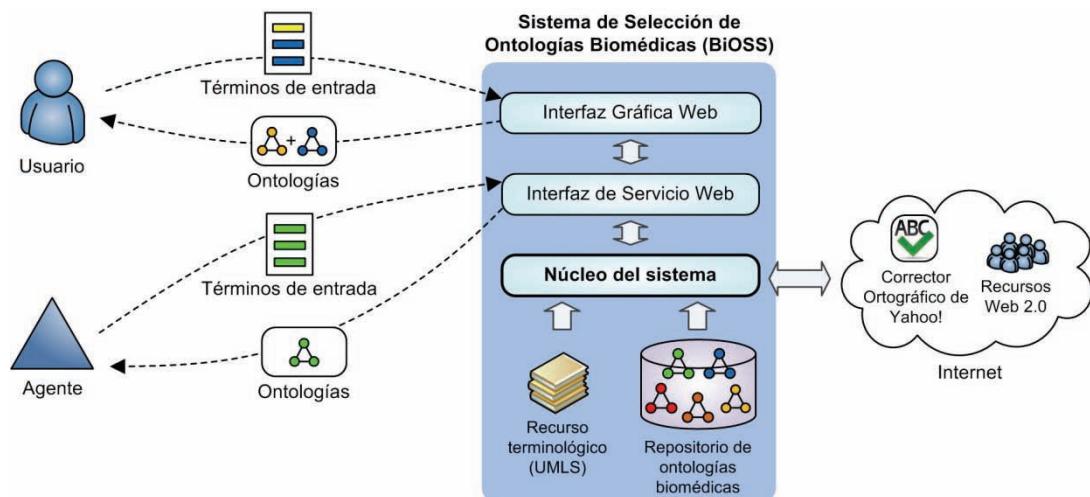


Figura 6.1. Arquitectura general del sistema de selección de ontologías biomédicas.

6.1.1 Estructura de paquetes

El diagrama de la figura 6.2 muestra la estructura de paquetes del sistema construido. A partir del paquete base `es.udc.imedir.java.bioss`, cuya nomenclatura se refiere a la información sobre país (España), institución (UDC), centro (IMEDIR), tecnología (Java) y aplicación (sistema BiOSS), cabe destacar los siguientes paquetes:

- `collectiveknowledgeresources`: contiene varias clases de acceso a los recursos de conocimiento colectivo (e.g. BioPortal, Wikipedia, etc.) que se utilizan para evaluar la popularidad de cada ontología.
- `evaluator`: este paquete contiene las clases que dirigen el proceso de evaluación de ontologías y el posterior proceso de selección.
- `math`: contiene varias operaciones matemáticas utilizadas por el sistema (e.g. cálculo de combinaciones sin repetición, utilizado para evaluar los resultados que proporcionan diferentes combinaciones de ontologías).

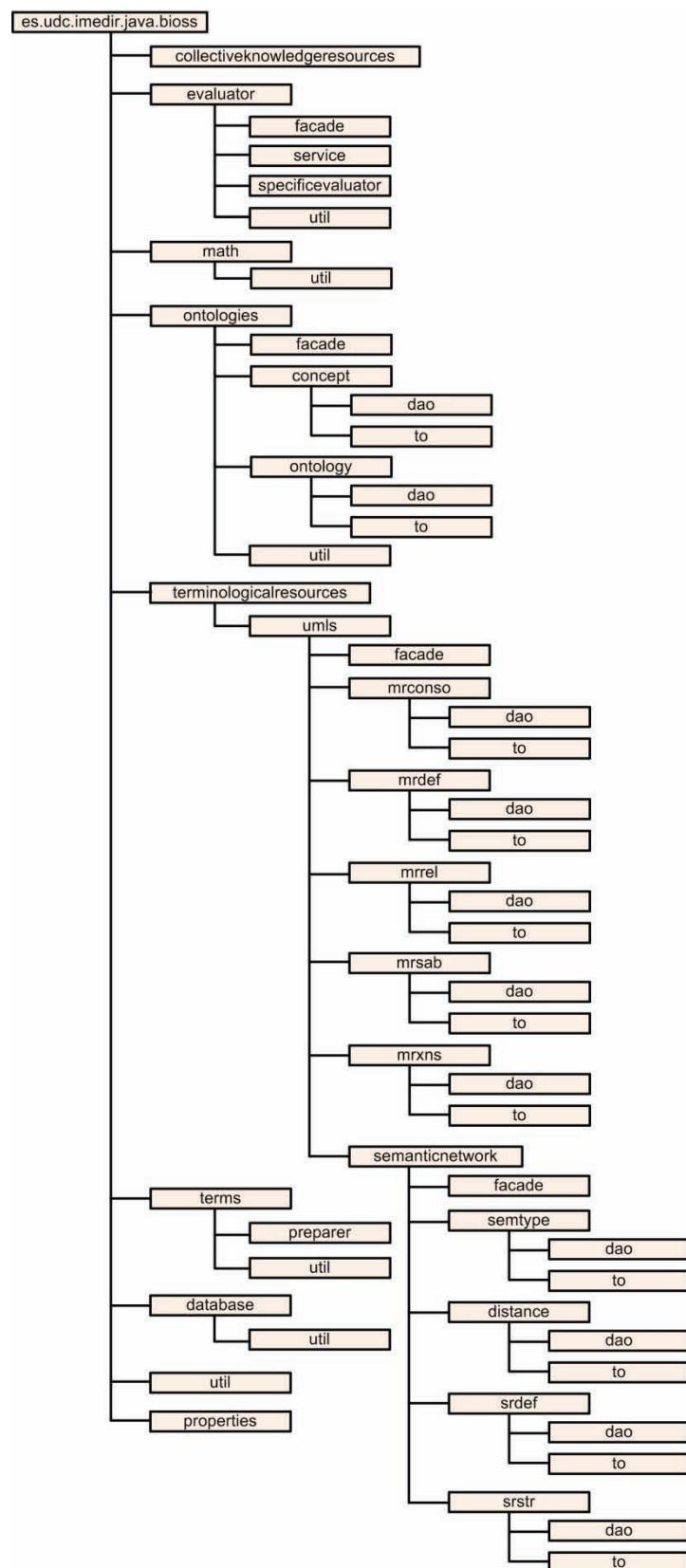


Figura 6.2. Estructura de paquetes del sistema de selección BiOSS.

- **ontologies**: abarca todas las operaciones relativas al manejo de las ontologías candidatas. Por ejemplo, contiene operaciones para almacenar una ontología procedente de otra fuente en el repositorio utilizado por el sistema, identificar correspondencias entre términos de entrada y conceptos de una ontología, recuperar toda la información asociada a un concepto determinado, etc.
- **terminologicalresources**: operaciones para el manejo de los recursos terminológicos en los que se basa el sistema (e.g. UMLS).
- **terms**: diversas operaciones terminológicas (e.g. normalización básica, normalización basada en UMLS, expansión semántica, corrección ortográfica, etc.) utilizadas a lo largo del proceso de selección.
- **database**: las clases de este paquete implementan la capa de acceso a datos. La implementación realizada es independiente del SGBD utilizado y optimiza el acceso a los datos mediante la implementación de un *pool* de conexiones.
- **util**: utilidades de propósito general (e.g. generador de ficheros de *log*).
- **properties**: paquete utilizado para gestionar de forma centralizada los parámetros de configuración del sistema (e.g. cadenas de conexión a base de datos).

También es importante destacar que, en el diseño de la aplicación, se han utilizado varios patrones arquitectónicos ampliamente conocidos (e.g. Model View Controller, Data Access Object, Transfer Object, Facade, etc.). El uso de estos patrones ha facilitado y agilizado el desarrollo del software, así como su futuro mantenimiento y potencial reutilización como parte de nuevos proyectos.

Tras explicar la estructura general del prototipo que se ha implementado, en los siguientes apartados se explicarán los detalles de implementación más relevantes. Para facilitar la comprensión, la explicación se realizará de forma ordenada, siguiendo la estructura del capítulo anterior, es decir, de acuerdo a las fases del proceso de selección de ontologías explicadas en el apartado 5.1.

6.2 Expansión semántica

En los siguientes apartados, se explica cómo se ha implementado el proceso de expansión semántica de las palabras clave de entrada, para ampliar cada una de ella con otros términos del mismo significado.

6.2.1 Normalización y corrección

Cada término de entrada se normaliza utilizando la herramienta de normalización de términos biomédicos proporcionada por UMLS (herramienta *norm*), perteneciente a las Herramientas Léxicas de UMLS. Se puede consultar información sobre esta herramienta en el apartado 2.1.5.2.3.

Para utilizar la herramienta de normalización, ha sido necesario instalar las Herramientas Léxicas de UMLS. Estas herramientas acompañan a cada distribución de UMLS, aunque también se pueden descargar de forma independiente⁵⁴. La instalación consiste en copiar varios ficheros con información sobre *stop words*, acentos diacríticos, conjunciones, etc., y en la ejecución de un *script* que carga los datos necesarios en una base de datos. También es necesario ajustar los parámetros de configuración en un fichero denominado *lvg.properties* (ver figura 6.3).

Una vez instaladas, se ha utilizado la API de las Herramientas Léxicas⁵⁵ para construir una clase Java capaz de invocar al normalizador (ver clase **UmlsNormalizer** de la figura 6.4). Esta clase dispone de un método **normalize**, que recibe como entrada un término y proporciona como salida el término normalizado de acuerdo a las Herramientas Léxicas de UMLS.

En cuanto a la corrección ortográfica, ésta se delega en el servicio de corrección ortográfica proporcionado por Yahoo⁵⁶. Este servicio recibe como entrada un término y proporciona como salida un nuevo término corregido ortográficamente, en caso de que el término inicial haya sido considerado erróneo. En la figura 6.5 se puede observar la salida proporcionada por el servicio de sugerencias ortográficas de Yahoo para el término *biomedicine*. Se puede observar que al término proporcionado como

⁵⁴ <http://lexsrv3.nlm.nih.gov/Specialist/Summary/lexicalTools.html/>

⁵⁵ <http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lvg/2010/docs/userDoc/api/api.html/>

⁵⁶ <http://developer.yahoo.com/search/web/V1/spellingSuggestion.html>

entrada le falta una “i”. El servicio detecta que el término no es correcto y devuelve como salida el término corregido (*biomedicine*). Para el acceso a este servicio se utiliza la clase **YahooSpellChecker** de la figura 6.5. Esta clase dispone de un método **check** que invoca el servicio de sugerencias ortográficas de Yahoo con el término que recibe como entrada, parsea la salida XML devuelta por el servicio para obtener la corrección y proporciona como salida el término corregido.

```
-----
# Directory and files
#
# LVG_DIR: the absolute path of the lexical tool directory
# LVG_STOP_WORD_FILE: the relative path (to LVG_DIR) of stop word file.
# LVG_NONINFO_WORD_FILE: the relative path (to LVG_DIR) of non-info word file.
# LVG_CONJ_WORD_FILE: the relative path (to LVG_DIR) of conjunction file.
# LVG_REMOVE_S_FILE: the relative path (to LVG_DIR) of removeS file.
#
# LVG_DIACRITICS_FILE: the relative path (to LVG_DIR) of diacritics file.
# LVG_LIGATURES_FILE: the relative path (to LVG_DIR) of ligature file.
# LVG_UNICODE_SYNONYM_FILE: the relative path (to LVG_DIR) of Unicode synonym file
# LVG_UNICODE_SYMBOLS_FILE: the relative path (to LVG_DIR) of symbols map file
# LVG_UNICODE_FILE: the relative path (to LVG_DIR) of Unicode map file
# LVG_NON_STRIP_MAP_UNICODE_FILE: the relative path (to LVG_DIR) of non-strip
Unicode map file
-----
LVG_DIR=C:/EjecucionTesis/lvg2009/
LVG_STOP_WORD_FILE=data/misc/stopWords.data
LVG_NONINFO_WORD_FILE=data/misc/nonInfoWords.data
LVG_CONJ_WORD_FILE=data/misc/conjunctionWord.data
LVG_REMOVE_S_FILE=data/misc/removeS.data
#
LVG_DIACRITICS_FILE=data/Unicode/diacriticMap.data
LVG_LIGATURES_FILE=data/Unicode/ligatureMap.data
LVG_UNICODE_SYNONYM_FILE=data/Unicode/synonymMap.data
LVG_UNICODE_SYMBOL_FILE=data/Unicode/symbolMap.data
LVG_UNICODE_FILE=data/Unicode/unicodeMap.data
LVG_NON_STRIP_MAP_UNICODE_FILE=data/Unicode/nonStripMap.data
-----
# Database and JDBC driver
#
# DB_TYPE: HSQLDB, MYSQL, or OTHER
# DB_DRIVER: the JDBC driver
# DB_HOST: Hostname of MySql database
# DB_NAME: database name for Lvg (default is lvg2009)
# DB_USERNAME: user name for using Lvg database (default is lvg)
# DB_PASSWORD: password for lvg user (default is lvg)
#
DB_TYPE=MYSQL
DB_DRIVER=org.hsqldb.jdbcDriver
DB_NAME=lvg2009
DB_HOST=193.147.41.219:3306
```

Figura 6.3. Fragmento del fichero lvg.properties, que contiene los parámetros de configuración de las Herramientas Léxicas de UMLS.

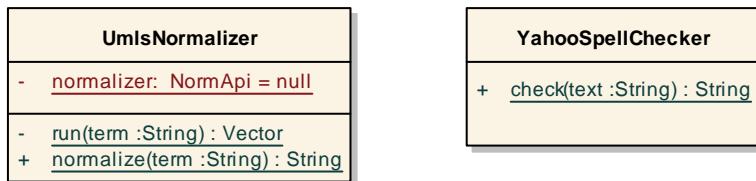


Figura 6.4. Detalle de las clases UmlsNormalizer y YahooSpellChecker.

```

<ResultSet xsi:schemaLocation="urn:yahoo:srch
http://api.search.yahoo.com/WebSearchService/V1/WebSearchSpellingResponse.xsd">

<Result>biomedicine</Result>

</ResultSet>

```

Figura 6.5. Sugerencia ortográfica proporcionada por el Servicio de Sugerencias Ortográficas de Yahoo para el término “biomedcine”.

6.2.2 Identificación de conceptos

En cuanto al proceso de identificación de conceptos, resulta interesante en este apartado de implementación explicar cómo se accede al recurso terminológico UMLS para obtener todos los significados de un término dado.

Cuando se dispone del término normalizado y, o, corregido, éste se busca en el índice de cadenas de texto normalizadas de UMLS. En la base de datos en la que UMLS se encuentra instalado, esta información se almacena en la tabla MRXNS_ENG (versión en inglés) y la búsqueda se realiza sobre el campo NSTR. Se trata de un campo tipo 'TEXT' que no tiene índice, por lo que las búsquedas son muy lentas. Por ello, ha sido necesario crear un índice para este campo, usando la opción "Set index column lenght" de MySQL, con un tamaño de campo de 255. En esta tabla, existirá una fila por cada significado del término normalizado.

En la tabla 6.1 se muestra un fragmento de la tabla MRXNS_ENG para el término normalizado *cavity*. La columna LAT indica el idioma (inglés), la columna NSTR contiene el término normalizado (*Normalized String*) y la columna CUI (*Concept Unique Identifier*) se refiere al identificador del concepto en UMLS correspondiente. De esta manera, para el término *cavity* se han identificado tres conceptos en UMLS (i.e. tres CUIs). A partir del CUI se puede obtener toda la información asociada al concepto (su nombre más frecuente, su tipo semántico, definición, sinónimos, etc.).

Tabla 6.1. Fragmento de la tabla mrxns_eng para el término *cavity*.

LAT	NSTR	CUI
ENG	cavity	C0011334
ENG	cavity	C0333343
ENG	cavity	C1510420

En la figura 6.6 se muestra un diagrama de secuencia para el proceso de normalización, corrección e identificación de conceptos.

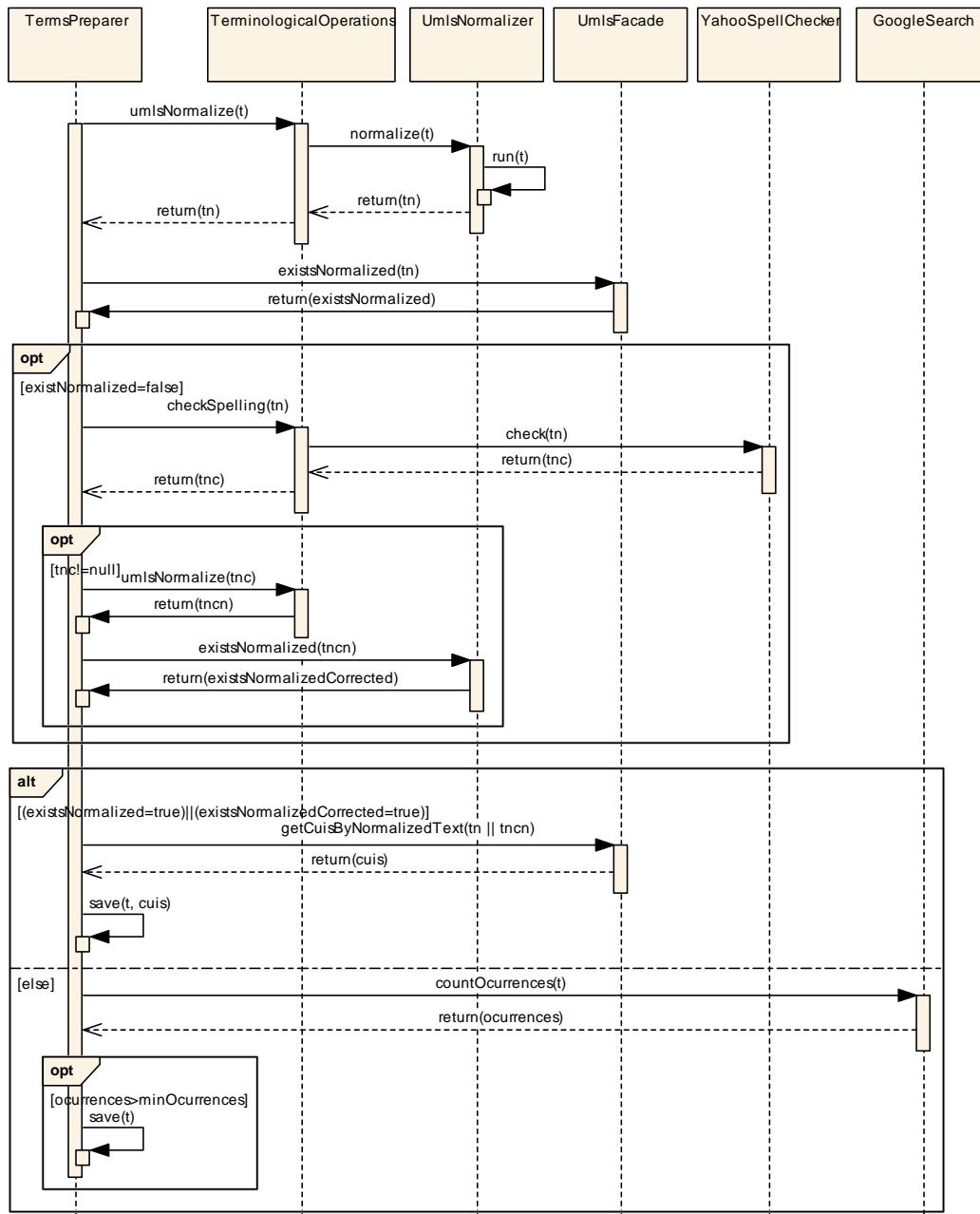


Figura 6.6. Diagrama de secuencia para el proceso de normalización, corrección e identificación de conceptos. Para simplificar el diagrama, se han obviado las llamadas a las clases que realizan el acceso a datos (i.e. DAOs).

6.2.3 Desambiguación

El proceso de desambiguación utiliza la información sobre los tipos semánticos de los conceptos de UMLS. Esta información se encuentra en la Semantic Network, que acompaña a la distribución de UMLS, y consiste en un conjunto de tablas de base de datos con información sobre los tipos semánticos de los conceptos del Metatesauro UMLS y las relaciones entre ellos. En el apartado 2.1.5.2.2 se proporciona más información sobre este recurso.

Para acceder a la Semantic Network y trabajar con la información que ésta proporciona, se ha implementado la clase `SemNetFacade`, perteneciente al paquete `es.udc.imedir.java.bioss.terminologicalresources.umls.semanticnetwork`. Esta clase se muestra en la figura 6.7, y proporciona diversas operaciones sobre la Semantic Network, como la obtención del camino mínimo entre dos nodos, la obtención del padre común más cercano (*lowest common ancestor*, LCA), la obtención del tipo semántico de un concepto cualquiera de UMLS, etc.

facade::SemNetFacade
<pre> - connectionStringSN: String - connectionStringSNAddData: String + SemNetFacade() + importSemTypesToMyDB(): void + saveSNDistancesToMyDB(): void + getRadaDistance(String, String): int + getRadaDistanceFromDB(String, String): int + getRadaDistancesFromDB(ArrayList<String>): int[] + getMinPathDistance(String, String): int + getMinPathDistances(ArrayList<String>): int[] + getMinPathDistanceFromDB(String, String): int + getMinPathDistancesFromDB(ArrayList<String>): int[] + getLCA(String, String): String + getLCA(ArrayList<String>): String + getPathToRoot(String): ArrayList<String> + getPathToRootLength(String): int + getPathToAncestor(String, String): ArrayList<String> + getPathToAncestorLength(String, String): int + getMinPathToNodeByIds(String, String): List<String> + getMinPathToNodeByNames(String, String): List<String> + disambiguate(Hashtable<String, ArrayList<String>>): Hashtable<String, String> + getMostRelatedSemanticType(String, ArrayList<String>): String + getPreferredSemType(Hashtable<String, Integer>): String + getSemTypesAndOccurrences(Hashtable<String, ArrayList<String>>): Hashtable<String, Integer> + getMCST(Hashtable<String, Integer>): ArrayList<String> + getGlobalDistances(Hashtable<String, Integer>, int[][]): Hashtable<String, Float> + getSemanticTypeNamesByConceptCuis(ArrayList<String>): ArrayList<String> + getSemanticTypeNameByConceptCui(String): String + getSTIdsBySTNames(ArrayList<String>): ArrayList<String> + getSTIdBySTName(String): String + getSTNameBySTId(String): String </pre>

Figura 6.7. Clase `SemNetFacade`.

Para calcular el tipo semántico preferido de un conjunto de tipos semánticos se utiliza la noción, ideada por el autor, de tipo semántico más central o MCST (*Most Central Semantic Type*), explicada en el apartado 5.2.3. El cálculo de MCST se realiza a partir de la longitud del camino mínimo entre dos tipos semánticos cualesquiera. Para el cómputo de este camino mínimo, se ha considerado la Semantic Network como un grafo no dirigido con todos los pesos de las aristas iguales, y se ha utilizado un algoritmo para el cálculo del camino mínimo en este tipo de grafo. El pseudocódigo del algoritmo utilizado se muestra en la figura 6.8.

```

caminoMinimo(nodoInicial, nodoFinal)
    nodosVisitados = vacío
    nodosAnteriores = vacío
    camino = vacío
    cola = vacío
    nodoActual = nodoInicial
    añadir nodoActual a cola
    añadir nodoActual a nodosVisitados

    mientras la cola no esté vacía hacer
        nodoActual = primer elemento de la cola
        eliminar nodoActual de la cola
        si nodoActual = nodoFinal entonces
            salir de bucle mientras
        fin si
        sino entonces
            nodosRelacionados =
                nodos directamente relacionados con nodoActual
            fin sino
            para cada nodo en nodosRelacionados hacer
                si nodosVisitados no contiene nodo entonces
                    añadir nodo a cola
                    añadir nodo a nodosVisitados
                    añadir nodoActual a nodosAnteriores de nodo
                fin si
            fin para
        fin mientras

        nodo = nodoFinal

        mientras nodo != null
            nodo = nodoAnterior de nodo
            añadir nodo a camino
        fin mientras

        ordenar camino inversamente

        devolver camino
    fin

```

Figura 6.8. Pseudocódigo del algoritmo para calcular el camino mínimo entre dos nodos de un grafo no dirigido con todos los pesos de las aristas iguales.

Por motivos de rendimiento y teniendo en cuenta que la Semantic Network de UMLS no varía de una ejecución a otra, se ha decidido calcular las distancias de los caminos mínimos desde cada tipo semántico a todos los demás, y almacenarlas en base de datos de tal manera que estas distancias estén disponibles en cualquier momento sin necesidad de ejecutar el algoritmo de cálculo de camino mínimo. Para esto, se ha creado un nuevo esquema de base de datos, que se mantiene independiente a las tablas de la Semantic Network, que consta de dos tablas, llamadas SEMTYPE y DISTANCE, las cuales almacenan los identificadores y nombres de cada tipo semántico y las distancias entre tipos semánticos, respectivamente. En la tabla 6.2 y tabla 6.3 se puede ver parte del contenido de las tablas SEMTYPE y DISTANCE.

Tabla 6.2. Fragmento de la tabla de base de datos SEMTYPE, que almacena los identificadores de los tipos semánticos (UI) y sus nombres (STY_RL).

UI	STY_RL
T003	Alga
...	...
T086	Nucleotide Sequence
T087	Amino Acid Sequence
T088	Carbohydrate Sequence
T089	Regulation or Law
T090	Occupation or Discipline
T091	Biomedical Occupation or Discipline
...	...

Tabla 6.3. Fragmento de la tabla de base de datos DISTANCE, que guarda las distancias de los caminos mínimos (DIST) para todos los posibles pares de tipos semánticos (UI1, UI2).

UI1	UI2	DIST
T003	T086	5
T003	T087	5
T003	T088	6
T003	T089	5
T003	T090	4
T003	T091	5
...

6.2.4 Expansión semántica

Tras finalizar el proceso de desambiguación, se dispone de un único significado asociado a cada término de entrada. En este punto, se realiza la expansión semántica del término propiamente dicha, que consiste en acceder al recurso terminológico (en este caso UMLS) y obtener todos los términos equivalentes a él (sinónimos).

La obtención de los sinónimos se realiza como se muestra en el diagrama de secuencia de la figura 6.9. A partir del identificador del concepto que se ha determinado para cada término (CUI), se obtienen todas las instancias del mismo en la tabla MRCONSO del Metatesauro UMLS. Cada una de estas instancias (filas de la tabla) contiene un término que representa al concepto. Todos estos términos se normalizan y se guardan, evitando almacenar sinónimos duplicados procedentes de diferentes ontologías de UMLS.

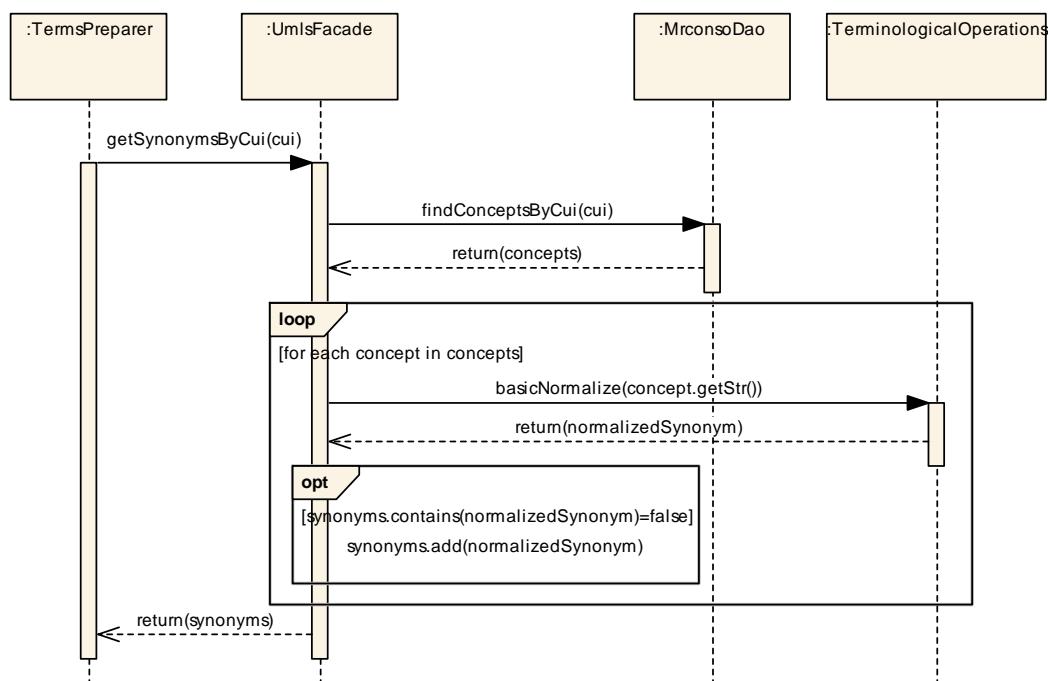


Figura 6.9. Diagrama de secuencia del proceso de expansión semántica de un concepto, cuyo identificador en UMLS es la variable *cui*.

A modo de ejemplo, en la tabla 6.4 se pueden ver los diferentes términos que proporcionan algunas de las ontologías del Metatesauro UMLS para el concepto cuyo CUI es *C0003483*, y que se suele representar por el término *aorta*.

Todos los términos que se pueden ver en la columna STR de la tabla son considerados por UMLS como sinónimos. Tras normalizarlos y eliminar los términos repetidos, se dispondría del conjunto de términos semánticamente expandidos para el concepto mencionado. De esta manera, el conjunto de sinónimos sería el siguiente: *{aorta, aortas, aortic, trunk of systemic arterial tree, trunk of aortic tree}*.

Tabla 6.4. Fragmento del contenido de la tabla MRCONSO del Metatesauro UMLS para el concepto cuyo CUI es C0003483. La columna SAB indica el código de la ontología que contiene al concepto. La columna STR muestra el término que representa al concepto en cada ontología.

CUI	SAB	...	STR	...
C0003483	LCH	...	<i>Aorta</i>	...
C0003483	MSH	...	<i>Aorta</i>	...
C0003483	MTH	...	<i>Aorta</i>	...
C0003483	UWDA	...	<i>Aorta</i>	...
C0003483	AOD	...	<i>aorta</i>	...
C0003483	CSP	...	<i>aorta</i>	...
C0003483	NCI	...	<i>Aorta</i>	...
C0003483	HL7V2.5	...	<i>Aorta</i>	...
C0003483	NCI	...	<i>aorta</i>	...
C0003483	LNC	...	<i>Aorta</i>	...
C0003483	FMA	...	<i>Aorta</i>	...
C0003483	LNC	...	<i>Aorta</i>	...
C0003483	MSH	...	<i>Aortas</i>	...
C0003483	NCI	...	<i>Aortic</i>	...
C0003483	UWDA	...	<i>Trunk of aortic tree</i>	...
C0003483	FMA	...	<i>Trunk of aortic tree</i>	...
C0003483	UWDA	...	<i>Trunk of systemic arterial tree</i>	...
C0003483	FMA	...	<i>Trunk of systemic arterial tree</i>	...

6.3 Recuperación de ontologías

Para poder llevar a cabo la selección de ontologías, es imprescindible disponer de un conjunto de ontologías candidatas al que poder acceder. En este apartado se explican los detalles de implementación del repositorio de ontologías biomédicas que se utilizará como base para llevar a cabo el proceso de selección. También se explica la forma de acceder a dicho repositorio para recuperar las ontologías.

6.3.1 Construcción de un repositorio de ontologías biomédicas

A pesar de que en la actualidad existen varios repositorios de ontologías a los que se puede acceder de forma pública (e.g. Swoogle, Watson) existen algunos inconvenientes que dificultan su reutilización en un sistema de selección como el propuesto:

1. **Carencia de ontologías biomédicas.** Son repositorios de propósito general, preparados para trabajar con los formatos de representación de ontologías más habituales (i.e. RDF y OWL), pero que no proporcionan soporte para formatos específicos de ontologías biomédicas (e.g. OBO). Debido a esto, estos repositorios no contienen la mayoría de las ontologías biomédicas más recientes. El repositorio de ontologías biomédicas de referencia es BioPortal, pero actualmente no proporciona servicios de acceso a las ontologías que contiene. Éste es el principal inconveniente que se ha encontrado al valorar la opción de utilizar un repositorio de terceros.
2. **Menor control.** El uso de un repositorio externo implica depender completamente de él. Posibles caídas del repositorio, cambios de configuración o de tecnología afectarán directamente a la disponibilidad del sistema de selección. Además, al utilizar un repositorio externo no es posible modificarlo para actualizar las versiones de las ontologías existentes, modificar los metadatos de las ontologías si son incorrectos, añadir nuevas ontologías, etc. Para asegurar un buen funcionamiento del sistema, interesa evitar esta dependencia.
3. **Menor rendimiento.** Acceder a un repositorio externo, ubicado remotamente, será en general más lento que acceder a un repositorio local. Además, el rendimiento también se puede ver afectado por problemas de red o debido a otros accesos al repositorio por parte de otros usuarios.

Debido a estas razones, se ha optado por construir un repositorio de ontologías biomédicas propio. Sin embargo, aunque los actuales repositorios públicos de ontologías son demasiado limitados para el desarrollo del sistema que se plantea, es necesario enfatizar la importancia que este tipo de repositorios tendrán en un ámbito

de Web Semántica en el futuro. A día de hoy, están apareciendo las primeras iniciativas que tratan de impulsar el desarrollo de estos repositorios (e.g. evento SERES2010⁵⁷) En varios años, existirán grandes repositorios semánticos que clasificarán ontologías en múltiples formatos de diferentes dominios, que se encontrarán actualizados, y que dispondrán de las herramientas necesarias para manejar las ontologías que contienen.

Para la implementación del repositorio, se han estudiado dos alternativas tecnológicas:

- a) **Uso de un *framework* específico para el almacenamiento y consulta de datos RDF** (e.g. Sesame⁵⁸, Virtuoso⁵⁹, etc.). Se trata de repositorios específicamente diseñados para el almacenamiento de ontologías en RDF o OWL, conocidos habitualmente como *Triple Stores*, ya que almacenan los datos en forma de ternas RDF: (sujeto-predicado-objeto). Permiten realizar consultas sobre ellas en formatos de consulta de ontologías estándar, como SPARQL, y disponen de varias opciones de implementación. Así, por ejemplo, un repositorio Sesame dispone de tres modos de almacenamiento: (1) Modo nativo, que consiste en un almacenamiento en disco propio de Sesame. (2) Memoria del sistema (éste es el modo de almacenamiento más rápido). (3) Base de datos, que consiste en almacenar las ontologías utilizando un SGBD existente (e.g. MySQL).
- b) **Uso de una base de datos a medida.** Esta opción consiste en diseñar un modelo de base de datos para almacenar los datos de interés de cada ontología, e implementar los mecanismos necesarios para acceder a dichos datos.

Tras estudiar detenidamente ambas opciones, se ha decidido utilizar un modelo de base de datos relacional diseñado específicamente para cubrir los requerimientos del sistema de selección. Los principales motivos que justifican esta decisión son los siguientes:

- **Pruebas realizadas.** Tras realizar varias pruebas con Sesame, se han observado dos problemas:

⁵⁷ <http://www.ontologydynamics.org/od/index.php/seres2010/>

⁵⁸ <http://www.openrdf.org/>

⁵⁹ <http://www.openlinksw.com/dataspace/%20dav/wiki/Main/VOSRDF/>

- Errores de carga. No ha sido posible cargar en el repositorio Sesame ontologías de gran tamaño (e.g. el NCI Thesaurus, cuyo fichero en formato OWL ocupa 190 MB aproximadamente).
 - Bajo rendimiento. Se han realizado varias pruebas con los modos de almacenamiento “nativo” y “memoria” de Sesame. En ambos casos, las consultas basadas en expresiones regulares (sentencia *regex* del lenguaje SPARQL) se ejecutan con demasiada lentitud para poder ser utilizadas en un sistema de selección como el que se propone. Estas consultas son necesarias para detectar correspondencias de tipo subcadena (e.g. *cell* en *white cell*) entre términos y ontologías. El rendimiento obtenido utilizando la sentencia LIKE de MySQL es muy superior.
- **Almacenamiento de datos adicionales.** Para agilizar el proceso de selección, resulta de interés almacenar diversa información que varía con poca frecuencia (e.g. número de referencias a cada ontología desde diferentes recursos Web) en lugar de calcularla en cada ejecución. Utilizando una base de datos diseñada “ad-hoc”, se puede contemplar el almacenamiento de esta información junto con los datos de cada ontología, sin necesidad de utilizar un almacenamiento adicional.
 - **Almacenamiento de ontologías de UMLS.** Algunas de las ontologías biomédicas más importantes se encuentran almacenadas en el Metatesauro UMLS, y no se dispone de los ficheros OWL o RDF (formato en ternas) de las mismas. Resulta más sencillo trasladar la información de estas ontologías directamente a otra base de datos diseñada para esto, que traducirla a un formato de triples para almacenarla en un repositorio de datos RDF.
 - **Flexibilidad.** En el caso de algunas ontologías ha sido necesario programar acciones “ad-hoc” para evitar peculiaridades en ciertos nombres de conceptos. Un ejemplo es la ontología Adverse Event Ontology, que contiene la subcadena *ae:* antes del nombre de cada concepto (e.g. *ae: connected temporal region*). En este caso ha sido necesario programar la eliminación de esta subcadena para evitar resultados erróneos al buscar términos en la ontología y almacenar los nombres de conceptos modificados en el repositorio. Un

repositorio “a medida” proporciona más flexibilidad a la hora de afrontar casos como éste.

Así, el repositorio de ontologías se ha implementado como una base de datos MySQL, cuyo diagrama se muestra en la figura 6.10. Esta base de datos contiene dos tablas: la tabla *ontology*, en la que se almacenan los datos de todas las ontologías del repositorio, y la tabla *concept*, con la información correspondiente a los conceptos de cada ontología.

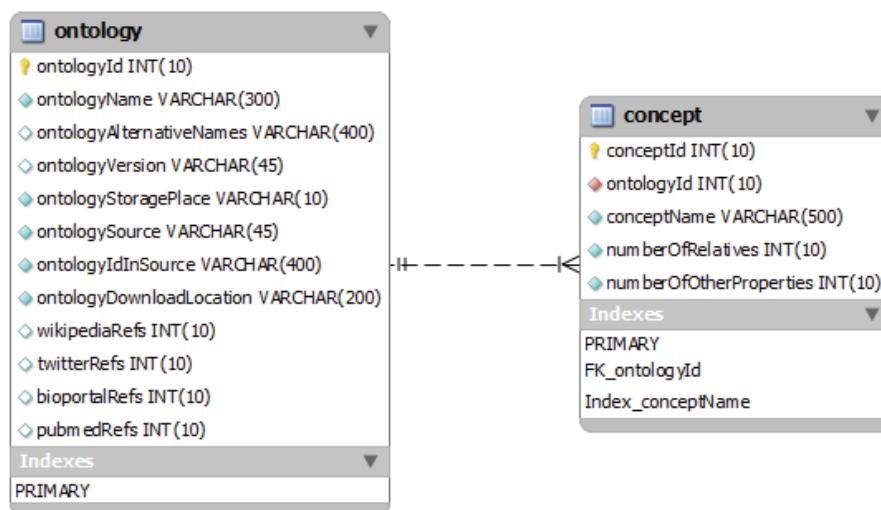


Figura 6.10. Diagrama de base de datos correspondiente al repositorio de ontologías.

La tabla *ontology* tiene los siguientes campos:

- **ontologyId:** identificador de la ontología (e.g. 124).
- **ontologyName:** nombre de la ontología (e.g. NCI Thesaurus).
- **ontologyAlternativeNames:** otros nombres frecuentes que recibe la ontología (e.g. National Cancer Institute Thesaurus).
- **ontologyVersion:** versión de la ontología (e.g. 2009AA)
- **ontologyStoragePlace:** lugar donde se almacenan los conceptos de la ontología. Actualmente, el valor de este campo para todas las ontologías de la BD es *mydb*, que indica que los conceptos están en la tabla *concept* de la BD.
- **ontologySource:** fuente de la que procede la ontología (e.g. UMLS, OBO Foundry, etc.).

- **ontologyIdInSource:** identificador de la ontología en la fuente de la que procede (e.g. NCI).
- **ontologyDownloadLocation:** sitio Web desde el que se ha obtenido la ontología (e.g. <http://www.nlm.nih.gov/research/umls/>).
- **wikipediaRefs:** referencias a la ontología desde el recurso Wikipedia (e.g. 24).
- **twitterRefs:** referencias a la ontología desde el recurso Twitter.
- **bioportalRefs:** referencias a la ontología desde BioPortal. El valor sólo puede ser 0 o 1, indicando si la ontología se encuentra en BioPortal o no.
- **pubmedRefs:** referencias a la ontología desde el recurso PubMed.

Y los campos de la tabla *concept* son los siguientes:

- **conceptId:** identificador del concepto (e.g. 3421).
- **ontologyId:** identificador de la ontología contenedora del concepto (e.g. 156).
- **conceptName:** término que representa al concepto (e.g. *abdominal pain*).
- **numberOfRelatives:** número de parientes del concepto (padres, hijos y hermanos).
- **numberOfOtherProperties:** número de propiedades del concepto. Aquí se tienen en cuenta las relaciones de la clase con otras clases, definiciones, sinónimos del nombre de la clase, restricciones sobre posibles valores o tipos de datos que puede adoptar la clase.

En la figura 6.11 y en la figura 6.12 se muestra un fragmento de la tabla *ontology* y de la tabla *concept*, respectivamente.

ontologyName	ontologyAlternat	ontologyVer	ontolo	onto	ontologyIdIn	ontologyDownloadLocation	wikipediaRefs	twitterRefs	bioportalRefs	pubmedRefs
Clinical Classifications Software	NULL	2005	mydb	umls	CCS	http://www.nlm.nih.gov/research/umls/	1	0	0	14
Computer-Stored Ambulatory Records	NULL	89-95	mydb	umls	COSTAR	http://www.nlm.nih.gov/research/umls/	0	0	0	0
CRISP Thesaurus	NULL	2006	mydb	umls	CSP	http://www.nlm.nih.gov/research/umls/	2	0	0	1
COSTART	NULL	1995	mydb	umls	CST	http://www.nlm.nih.gov/research/umls/	10	7	1	17
DXplain	NULL	1994	mydb	umls	DXP	http://www.nlm.nih.gov/research/umls/	3	0	0	20
Foundational Model of Anatomy	NULL	2_0	mydb	umls	FMA	http://www.nlm.nih.gov/research/umls/	6	0	1	0
Gene Ontology	NULL	2008_04...	mydb	umls	GO	http://www.nlm.nih.gov/research/umls/	81	11	1	2950
Healthcare Common Procedure Coding System	NULL	2009	mydb	umls	HCPGS	http://www.nlm.nih.gov/research/umls/	9	1	0	63
HL7 Vocabulary Version 2.5	NULL	2003_08...	mydb	umls	HL7V2.5	http://www.nlm.nih.gov/research/umls/	0	0	0	0
HL7 Vocabulary Version 3.0	NULL	2006_05	mydb	umls	HL7V3.0	http://www.nlm.nih.gov/research/umls/	0	0	0	0
HUGO Gene Nomenclature	NULL	2008_03	mydb	umls	HUGO	http://www.nlm.nih.gov/research/umls/	23	0	0	35
ICD-10-PCS	NULL	2008	mydb	umls	ICD10PCS	http://www.nlm.nih.gov/research/umls/	10	0	1	22
ICD-9-CM	NULL	2009	mydb	umls	ICD9CM	http://www.nlm.nih.gov/research/umls/	37	16	0	9178
International Classification of Primary Care	NULL	1993	mydb	umls	ICPC	http://www.nlm.nih.gov/research/umls/	24	0	1	145
Library of Congress Subject Headings	NULL	1990	mydb	umls	LCH	http://www.nlm.nih.gov/research/umls/	23	0	0	4

Figura 6.11. Fragmento de la tabla *ontology*.

conceptId	ontologyId	conceptName	numberOfRelatives	numberOfOtherProperties
623220	118	mental disorders	28	0
623221	118	benign neoplasm of colon	5	0
623222	118	benign neoplasm of ovary	5	0
623223	118	biliary tract disease	18	2
623224	118	bipolar affective disorder	6	0
623225	118	birth trauma	7	2
623226	118	cancer of bladder	3	2
623227	118	bladder neck obstruction	2	0
623228	118	coagulation defects	3	0
623229	118	blood transfusion	42	2
623230	118	bone marrow biopsy	4	2
623231	118	bone marrow transplant	4	2
623232	118	concussion	2	0
623233	118	cancer of breast	16	2

Figura 6.12. Fragmento de la tabla *concept* en la que se muestran algunos conceptos de la ontología Clinical Classifications Software (*ontologyId* = 118).

6.3.2 Almacenamiento y acceso a las ontologías

Una vez diseñado y construido el repositorio de ontologías, el siguiente paso fue almacenar en él un conjunto de ontologías biomédicas. El almacenamiento de las ontologías en el repositorio y el posterior acceso al repositorio para consultar las ontologías almacenadas se ha realizado utilizando las clases del paquete *es.udc.imedir.java.bioss.ontologies*, cuyo detalle se muestra en la figura 6.13.

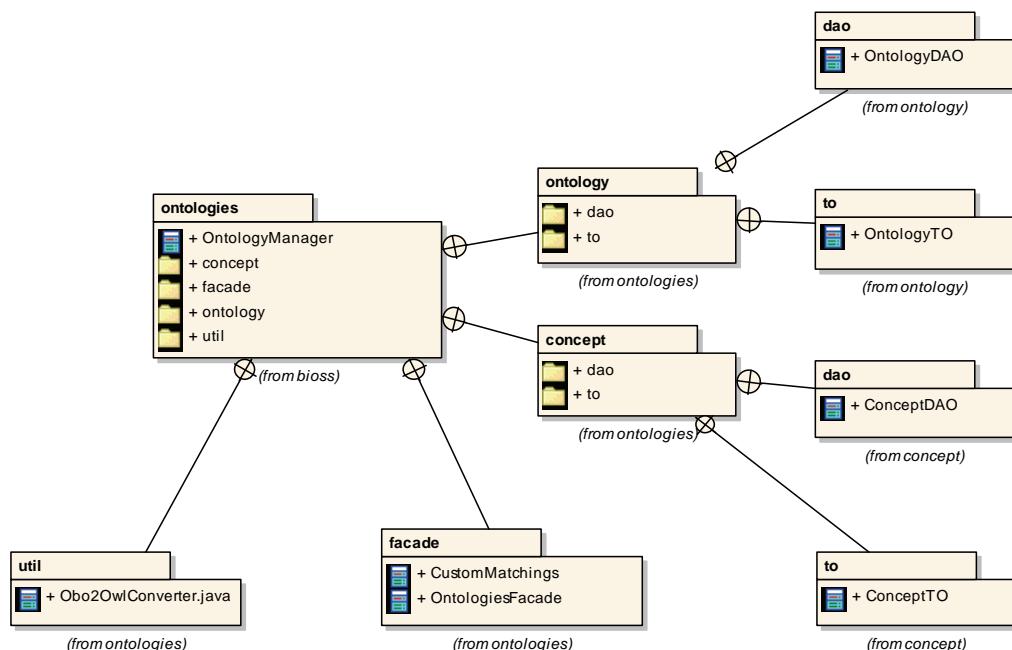


Figura 6.13. Detalle del paquete *es.udc.imedir.java.bioss.ontologies*.

Las principales operaciones que proporciona este paquete se encuentran en la clase `OntologiesFacade` (ver figura 6.14). Contiene operaciones para el almacenamiento de ontologías en el repositorio, para actualizar el número de referencias a cada ontología desde recursos Web, para obtener las correspondencias (*mappings* o *matchings*) entre un término y una ontología, etc.

facade::OntologiesFacade
<pre>+ getOntologyNames() : String[] + getAllNamesForOntology(long) : ArrayList<String> + getOntologyIds() : Long[] + getOntologyName(Long) : String - addOnlyOntologyToMyDB(OntologyTO) : OntologyTO + calculateNumberOfReferences(long) : int[] + calculateNumberOfReferences(ArrayList<String>) : int[] - addOntologyAndConceptsFromFileToMyDB(String, String, String, String, String, String, String) : void + addOntologiesAndConceptsFromUmlsToMyDB() : void + addOntologiesAndConceptsFromFilesToMyDB() : void + findOntologyById(Long) : OntologyTO + findOntologyByName(String) : OntologyTO + getOntologies() : ArrayList<OntologyTO> + getAllConceptsFromOntology(long) : ArrayList<ConceptTO> + getAllConceptsFromUMLSOntology(long, String) : ArrayList<ConceptTO> + getMatchingForConcept(Long, ArrayList<String>) : CustomMatchings + getMatchingsForTerm(Long, String, int) : CustomMatchings - getMatchingsForTermInMyDb(Long, String, int) : CustomMatchings - getMatchingsForTermInUmls(String, Long, String, int) : CustomMatchings + updateReferencesFromWebResources() : void + getRandomConcepts(int) : ArrayList<ConceptTO></pre>

Figura 6.14. Operaciones de la clase *OntologiesFacade*.

En cuanto al almacenamiento de las ontologías en el repositorio, se han contemplado dos casos:

- **Ontologías de UMLS.** Las ontologías pertenecientes al Metatesauro UMLS se encuentran almacenadas en una base de datos con una estructura a medida (se puede ver el detalle en la figura 2.12). Debido a esto, trasladar estas ontologías al repositorio ha requerido implementar operaciones específicas que consisten en acceder a la base de datos local en la que se ha instalado el Metatesauro UMLS para obtener los datos de la ontología en cuestión y sus conceptos, y almacenarlos en el repositorio local. Se han incorporado al repositorio las ontologías de la versión 2009AA de UMLS.

- **Ontologías en formato OWL o RDF.** En el caso de ontologías en estos formatos, se ha utilizado la API Jena⁶⁰ para acceder a la información de las ontologías.
- **Ontologías en formato OBO.** Muchas de las ontologías biomédicas existentes en la actualidad se encuentran almacenadas en el formato OBO, que es un formato específicamente diseñado para la representación de este tipo de ontologías. Para este caso, se ha creado la clase `Obo2OwlConverter.java`, del paquete `es.udc.imedir.java.bioss.ontologies.util`, que permite realizar la conversión de una ontología en formato OBO a formato OWL y así poder manejarla mediante la API Jena. La clase desarrollada se apoya en una conocida API para la manipulación de ficheros en formato OWL, conocida como la OWL API⁶¹.

Aunque una ontología se encuentre expresada en un formato estándar como OWL, muchas veces carece de datos esenciales (e.g. versión, o incluso el nombre), o estos datos no siempre se expresan de la misma forma. Esto impide el desarrollo de un método automático que identifique y extraiga dichos metadatos de cada ontología para su incorporación al repositorio. Debido a esto, se ha elaborado un fichero en formato XML (ver figura 6.15), que contiene la información necesaria para comenzar el procesamiento de cada ontología en formato OBO, OWL o RDF. Este fichero se procesa de forma automática para trasladar todos los datos de la ontología y sus conceptos al repositorio de ontologías.

En el caso de las ontologías procedentes de UMLS, el Metatesauro contiene toda la información requerida por el repositorio siguiendo una estructura clara. En este caso, los datos han podido extraerse automáticamente del Metatesauro e incorporarlos al repositorio, sin necesidad de crear un fichero intermedio.

El repositorio de ontologías que se ha creado, contiene actualmente 200 ontologías del ámbito biomédico, con un total de 1.860.881 conceptos, procedentes de fuentes como UMLS, BioPortal o la iniciativa OBO Foundry. La lista completa de ontologías contenidas en el repositorio se puede ver en el anexo I. Una vez almacenadas, las

⁶⁰ <http://jena.sourceforge.net/>

⁶¹ <http://owlapi.sourceforge.net/>

ontologías se consultan utilizando los diferentes métodos proporcionados por la fachada `OntologiesFacade`, anteriormente mencionada (ver figura 6.14).

```
<...>
<ontology>
    <source>OBO Foundry</source>
    <path>C:/ontologies/bioportal/deOBO/human-dev-anat-timed.owl</path>
    <name>Human developmental anatomy, timed version</name>
    <alternatenames>Human developmental anatomy ontology</alternatenames>
    <version>1.3</version>
    <idinsource></idinsource>
    <webpage>http://genex.hgu.mrc.ac.uk/</webpage>
</ontology>
<ontology>
    <source>OBO Foundry</source>
    <path>C:/ontologies/bioportal/deOBO/human_phenotype.owl</path>
    <name>Human Phenotype Ontology</name>
    <alternatenames></alternatenames>
    <version>13/11/2010</version>
    <idinsource></idinsource>
    <webpage>http://www.human-phenotype-ontology.org/</webpage>
</ontology>
<...>
```

Figura 6.15. Fragmento del fichero `ontologies.xml`. Este fichero contiene la información necesaria para trasladar cada ontología en formato OBO, OWL o RDF al repositorio.

6.4 Evaluación de ontologías

El paso clave del proceso de selección de ontologías es la evaluación de ontologías, que consiste en medir cómo de adecuada es cada ontología del repositorio de acuerdo a los términos de entrada proporcionados al sistema. Como se ha explicado en el capítulo anterior, la evaluación de cada ontología se realiza de acuerdo a tres criterios: (1) Evaluación de la cobertura del contexto. (2) Evaluación de la riqueza semántica. (3) Evaluación de la popularidad.

En la figura 6.16 se presenta el contenido del paquete `es.udc.imedir.java.bioss.evaluator`, que contiene las clases con la lógica del proceso de evaluación y selección de ontologías. Las operaciones que controlan el proceso de evaluación se agrupan en la clase `EvaluatorFacade` del subpaquete `facade` (ver figura 6.17). Esta clase dispone de operaciones que permiten calcular los términos expandidos para los términos de entrada, evaluar las ontologías de acuerdo a

los tres criterios establecidos, agregar los resultados en un único valor y seleccionar la ontología u ontologías más adecuadas de acuerdo a los resultados de la evaluación. En la figura 6.18 se muestra un diagrama que muestra el proceso de selección de ontologías, de forma general.

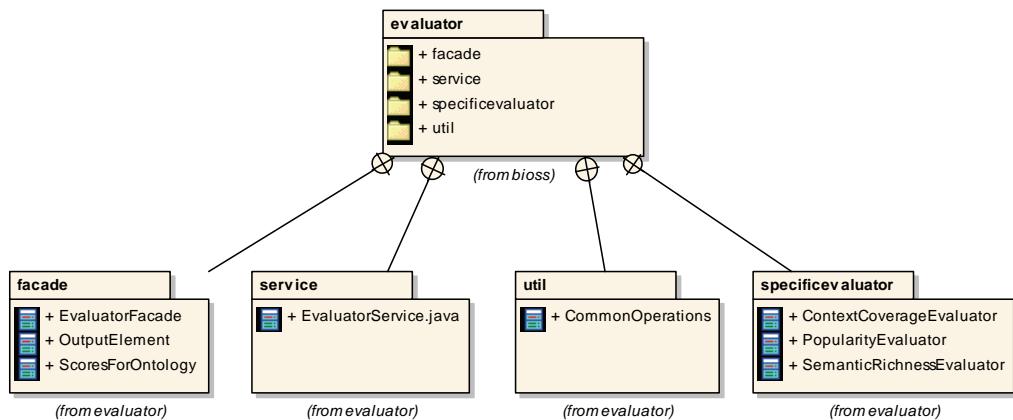


Figura 6.16. Detalle del paquete `es.udc.imedir.java.bioss.evaluator`.

6.4.1 Evaluación de la cobertura del contexto

La evaluación de la cobertura del contexto pretende medir en qué grado la ontología cubre las palabras clave que suponen la entrada del sistema de selección. Debido a esto, el punto clave de esta fase de evaluación es identificar cuántos de los conceptos subyacentes a dichas palabras clave (que han sido identificados durante las fases de identificación de conceptos y desambiguación) están contenidos en la ontología. Para esto, se contabilizan el número de términos iniciales que se encuentran contenidos en la ontología, considerando que un término está en la ontología si contiene un concepto cuyo nombre se corresponde con el del término, o con un término equivalente al mismo (i.e. alguno de sus sinónimos, obtenidos durante la fase de expansión semántica).

El proceso de cálculo de los conceptos de una ontología que cubren un conjunto de términos de entrada, representados cada uno de ellos por un conjunto de sinónimos, se puede ver en el diagrama de secuencia de la figura 6.19.

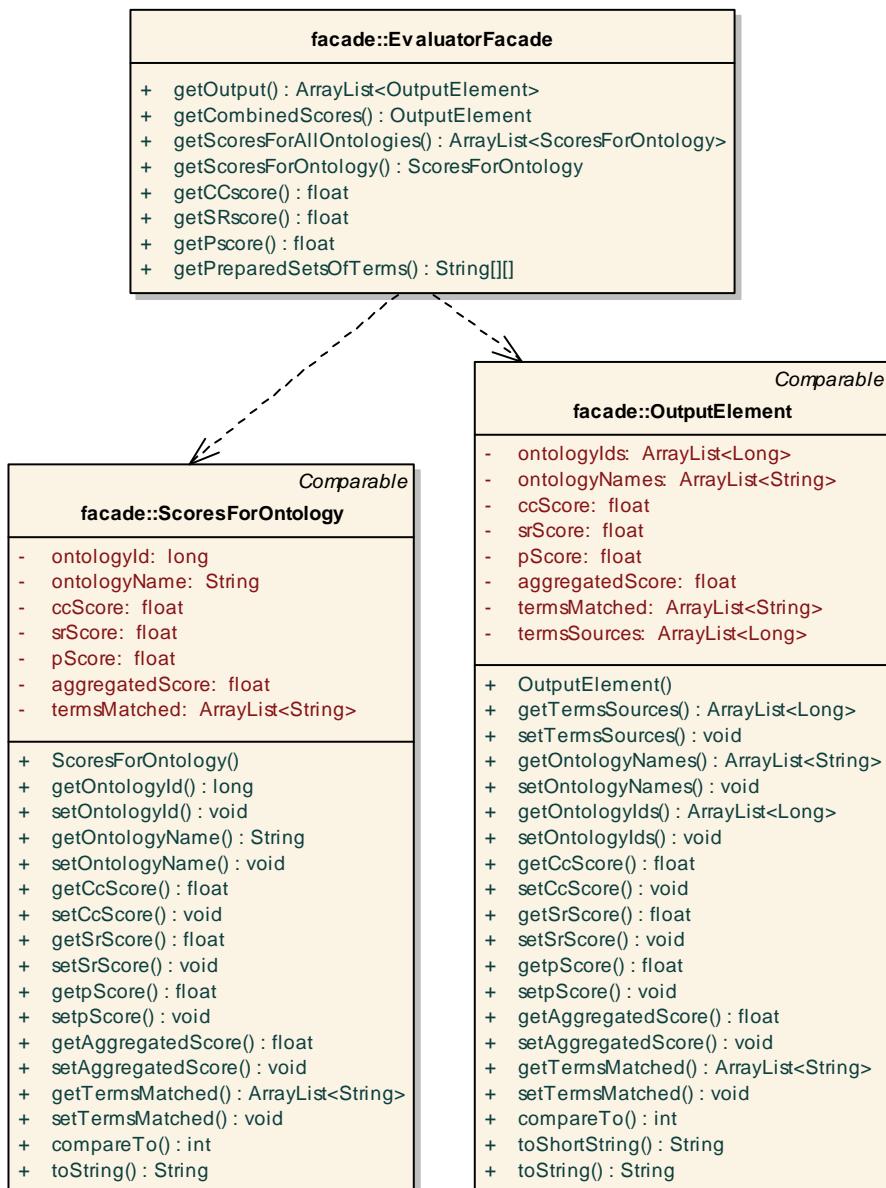


Figura 6.17. Clases del paquete *es.udc.imedir.java.bioss.evaluator.facade*. Para simplificar el diagrama, se ha obviado la representación de los parámetros de entrada de las operaciones.

La clase ConceptDao es la que realiza el acceso a la BD para comprobar si la ontología contiene un término determinado. Esta comparación se trata de un *matching* exacto, es decir, determinará que un término está en la ontología si coincide exactamente con el nombre de uno de los conceptos de la misma. Antes de esta comparación, el término se normaliza para evitar resultados incorrectos debido a dobles espacios, acentos, etc. Los nombres de los conceptos en el repositorio de ontologías ya se encuentran normalizados, pues se normalizan durante el proceso de

almacenamiento en el repositorio. Además, esta comparación también se realiza eliminando todos los espacios en blanco, de tal manera que, por ejemplo, los términos *white cell* y *whitecell* se considerarán equivalentes. Existen muchas ontologías biomédicas para las que se han eliminado los espacios en blanco en los nombres de conceptos, por lo que este tipo de comparación es imprescindible para obtener buenos resultados. Para realizar la búsqueda de conceptos cuyo nombre es igual que el de un término determinado, se utiliza la consulta MySQL que se puede ver en la figura 6.20.

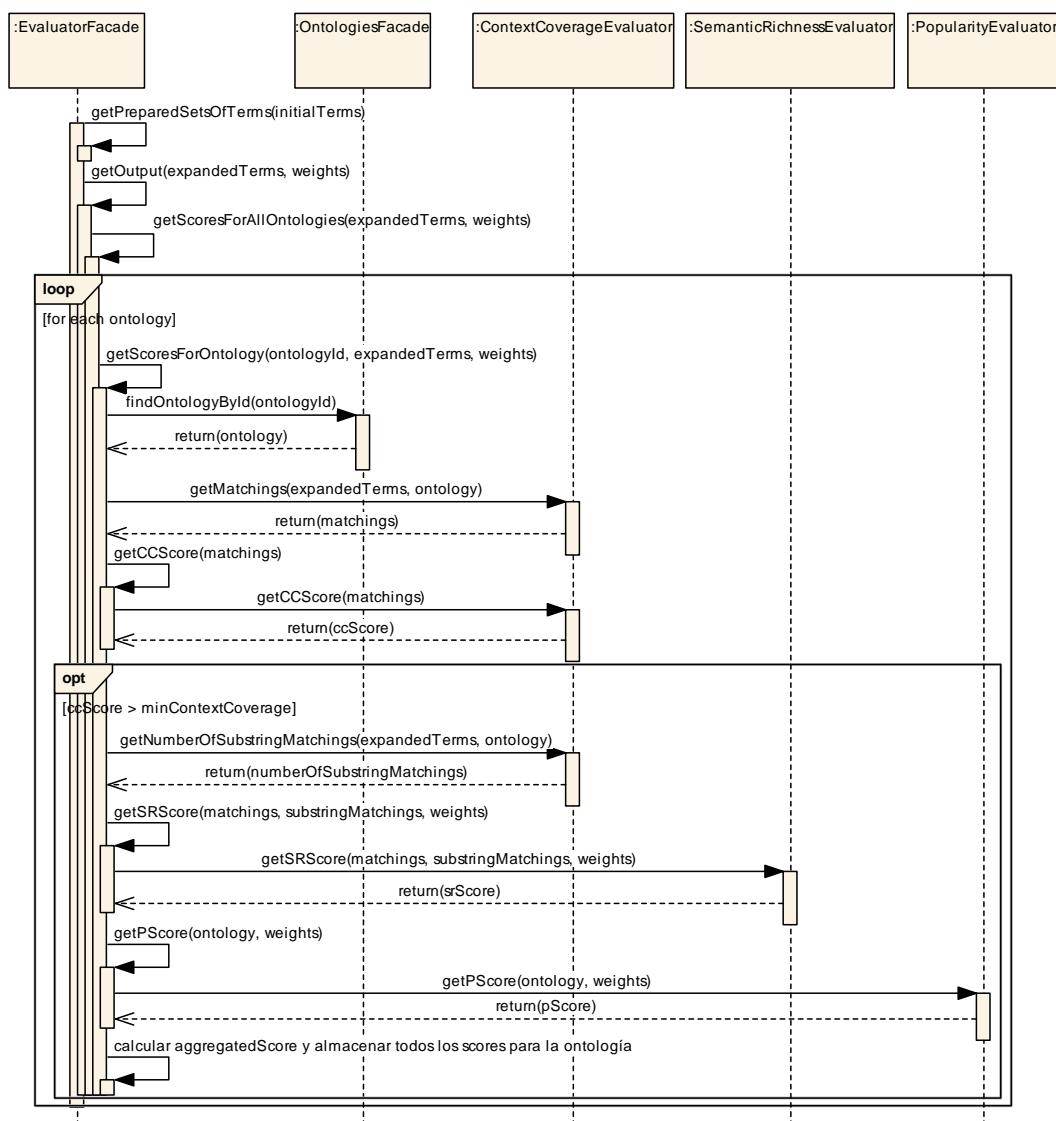


Figura 6.18. Diagrama de secuencia que muestra el proceso de evaluación y selección de ontologías.

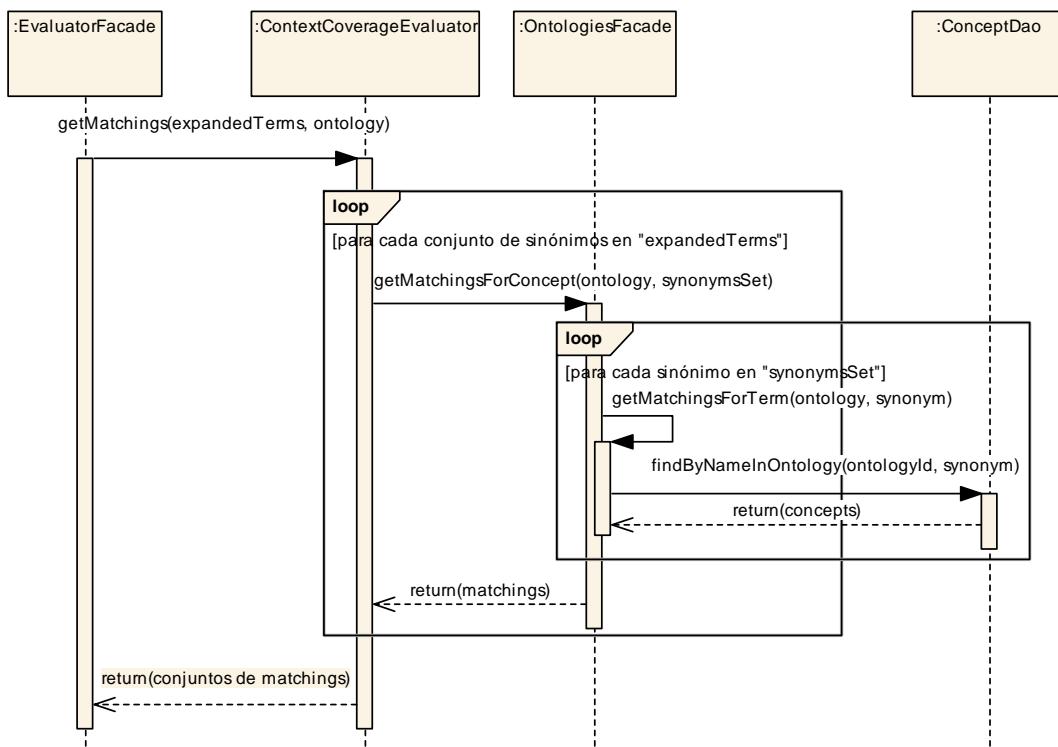


Figura 6.19. Diagrama de secuencia para el proceso de cálculo de correspondencias entre los términos de entrada (representados cada uno de ellos por un conjunto de sinónimos) y una ontología determinada. La variable *expandedTerms* contiene varios conjuntos de sinónimos, uno por cada término de entrada.

```

SELECT * FROM concept
WHERE (ontologyId = A) AND
((conceptName LIKE (t)) OR (conceptName LIKE (tw)))
  
```

Figura 6.20. Consulta MySQL utilizada para obtener todos los conceptos de la ontología de identificador A cuyo nombre es igual al término *t*. La variable *tw* se refiere al término *t* tras eliminar los espacios en blanco.

6.4.2 Evaluación de la riqueza semántica

La evaluación de la riqueza semántica consiste en calcular una medida del nivel de detalle que proporcionan los conceptos de una ontología que cubren los términos de entrada. Esta evaluación se realiza de acuerdo a tres aspectos diferentes: (1) **Parientes** directos de cada concepto (padres, hijos y hermanos). (2) **Información adicional** que la ontología contiene para cada concepto (e.g. definiciones, relaciones con otros conceptos, comentarios, etc.). (3) **Conocimiento similar** al concepto, es decir, otros conceptos de la ontología con significado parecido a cada concepto dado. En los siguientes apartados, se explica cómo se ha realizado la evaluación de estos tres

aspectos. Finalmente, se explica cómo se utilizan estas tres puntuaciones para calcular la medida de riqueza semántica de una ontología.

6.4.2.1 Evaluación de parentesco

Este tipo de evaluación requiere disponer del número de parientes directos (padres, hijos y hermanos) de cualquier concepto de una ontología. Para esto, durante el proceso de almacenamiento de una ontología en el repositorio de ontologías, se calcula el número de parientes para cada concepto, y este número se almacena en el campo `numberOfRelatives` de la tabla `concept`, en la base de datos correspondiente al repositorio de ontologías (los detalles de esta base de datos se han explicado en el apartado 6.3.1). Al disponer de este dato precalculado, se evita tener que volver a computar el número de parientes de cada concepto en cada ejecución y se agiliza el proceso de evaluación.

En el caso de que la ontología que se pretende almacenar en el repositorio se encuentre en formato OWL o RDF, el número de parientes de cada concepto se contabiliza utilizando el método `countClassRelatives`, de la clase `OntologyManager` (ver figura 6.21), que se encuentra en el paquete `es.udc.imedir.java.bioss.ontologies`. Este método usa las operaciones proporcionadas por la API Jena para contar el número de padres, hijos y hermanos de cada concepto. Si la ontología se encuentra en formato OBO, ésta sufre un proceso de traducción a OWL y luego se procesa como se ha explicado.

ontologies::OntologyManager
<ul style="list-style-type: none"> + <code>getLocalName(OntClass) : String</code> + <code>getLabel(OntClass, String) : String</code> + <code>getConceptsFromOntologyInDisk(String, long, String) : ArrayList<ConceptTO></code> + <code>countClassRelatives(OntClass) : int</code> + <code>countClassOtherProperties(OntClass) : int</code> + <code>createConcepts(ArrayList<ConceptTO>, int) : void</code> + <code>createConcept(ConceptTO) : ConceptTO</code>

Figura 6.21. Operaciones de la clase `OntologyManager`.

facade::UMLSFacade
<pre>+ getNumberOfOtherProperties(String, String) : int + getNumberOfRelativesForConceptInSource(String, String) : int + getNumberOfOtherRelationsForConceptInSource(String, String) : int + getDefinitionsByCuiAndSource(String, String) : ArrayList<MrdefTO> + areSynonyms(String, String) : boolean + exists(String) : boolean + existsNormalized(String) : boolean + exist(String, String) : boolean + findByNameInOntology(long, String, String, int) : ArrayList<ConceptTO> + getAllConceptsFromOntology(String) : ArrayList<MrconsTO> + getSynonymsByCuis(ArrayList<String>) : ArrayList<ArrayList<String>> + getSynonymsByCui(String) : ArrayList<String> + getSynonyms(String) : ArrayList<String> + getPreferredNameByCui(String) : String + getDefinitionByCui(String) : String + getOntologies() : ArrayList<OntologyTO> + getCuisByText(String) : ArrayList<String> + getCuisByNormalizedText(String) : ArrayList<String></pre>

Figura 6.22. Operaciones de la clase *UMLSFacade*.

En el caso de las ontologías procedentes de UMLS, los parientes de cada concepto se calculan utilizando la operación `getNumberOfRelativesForConceptsInSource` de la clase `UMLSFacade` (perteneciente al paquete `es.udc.imedir.java.bioss.terminologicalresources.uml.s.facade`). Esta clase accede a la base de datos en la que se encuentra almacenada el conjunto de ontologías de UMLS (Metatesauro UMLS) y cuenta el número de parientes de cada concepto, usando para ello las relaciones jerarquía de UMLS.

6.4.2.2 Evaluación de información adicional

Este proceso se realiza en base a la cantidad de información que una ontología proporciona para un concepto. Se tiene en cuenta cualquier información adicional, exceptuando los aspectos ya tenidos en cuenta en otras fases de la evaluación (i.e. información proporcionada por las relaciones `is_a`, ya considerada durante la evaluación de parentesco; el nombre del concepto; y la información sobre las instancias de cada concepto, en caso de existir). De esta manera, esta información adicional comprendería las relaciones del concepto con otros conceptos, sus definiciones, sinónimos, restricciones, etc. De la misma forma que en el caso de la evaluación de parentesco, este cálculo se realiza durante el proceso de almacenamiento de cada ontología en el repositorio. La cantidad de elementos que proporcionan información adicional de un concepto, se almacenan en el campo `numberOfOtherProperties` de

la tabla *concept*, (cuyos detalles se pueden encontrar en el apartado 6.3.1). Dependiendo de si la ontología original se encuentra en un formato estándar (e.g. OWL) o procede de UMLS, la forma de calcular los elementos de información adicional varía.

Si la ontología se encuentra en formato OWL o RDF, el cálculo consiste en contar el número de propiedades de la clase, exceptuando las propiedades `subClassOf` (relación subclase), `type` (tipo de la clase), y `label` (nombre de la clase). Esto se realiza a través del método `countClassOtherProperties` (ver figura 6.21), que utiliza la API Jena para interpretar el fichero OWL o RDF de la ontología.

Si la ontología se encuentra originalmente en el Metatesauro UMLS, la forma de proceder consiste en acceder a la base de datos en la que éste se encuentra almacenado (en el apartado 2.1.5.2.1 se pueden ver las tablas que componen la BD del Metatesauro UMLS y sus atributos) y obtener el número de elementos de información adicional. Este número se calcula como la suma de:

1. El número de definiciones que una ontología proporciona para un concepto. Para obtenerlo, se accede a la tabla MRDEF de la base de datos.
2. El número de relaciones que tiene el concepto. Para esto se accede a la tabla MRREL, y se tienen en cuenta únicamente los tipos de relación de sinonimia (código SY) y cualquier otro tipo de relación que no sea de parentesco (código RO).

6.4.2.3 Evaluación de conocimiento similar

En este paso de evaluación, se trata de medir cuánto conocimiento posee la ontología, en el ámbito del contexto en el que está siendo evaluada. Para esto, dado un determinado concepto de la ontología, se cuentan el número de conceptos de la ontología cuyo significado es similar al significado del concepto.

De forma más específica, esto se consigue contando los conceptos de la ontología cuyo nombre contiene al nombre del concepto que se está evaluando, o al nombre de alguno de los sinónimos de este concepto. Para esto, según se ha explicado en el apartado 5.4.2.3, la clave está en contar el número de correspondencias de subcadena (*substring matchings*) en la ontología, para el nombre del concepto dado y todos sus sinónimos.

Como en el repositorio se dispone de los nombres de todos los conceptos de las ontologías, en un formato normalizado (i.e. sin dobles espacios, guiones bajos, caracteres diacríticos, etc.), buscar los conceptos similares a un concepto dado consiste en consultar la tabla *concept* de la BD MySQL en la que se encuentra el repositorio (los detalles de esta BD se encuentran en el apartado 6.3.1) y contar todos los conceptos cuyo campo *conceptName* contiene al nombre del concepto, o al nombre de alguno de los sinónimos del concepto (ver figura 6.23).

```
SELECT * FROM concept
WHERE (ontologyId = <id>) AND
((conceptName LIKE ('%<term>%')) AND
(conceptName REGEXP ('(^|[^[:alpha:]])<term>($|[^[:alpha:]]')'))))
```

Figura 6.23. Consulta utilizada para obtener las correspondencias de subcadena. Se asume que *<id>* es el identificador de la ontología sobre la que se realiza la búsqueda y que *<term>* es el término de búsqueda. El campo de la base de datos en el que se almacenan los nombres de los conceptos de las ontologías es *conceptName*.

Esto se realiza utilizando el método *getNumberOfSubstringMatchings* de la clase *ContextCoverageEvaluator*, que combina las funciones LIKE y REGEXP de MySQL para encontrar los conceptos que contienen una cadena de texto determinada tratando de obtener el mejor rendimiento posible. Las restricciones impuestas por LIKE y REGEXP se solapan, pero proporcionan un resultado de forma mucho más rápida que usando únicamente REGEXP. La función LIKE, mucho más rápida, realiza un primer filtrado, encontrando todas las correspondencias de subcadena, sea cual sea el carácter anterior o posterior al término buscado. REGEXP realiza el segundo filtrado sobre el resultado obtenido por LIKE, quedándose únicamente con aquellos resultados en los que el carácter anterior y posterior al término de búsqueda sea un carácter no alfanumérico o bien el carácter de inicio o fin de cadena, evitando correspondencias incorrectas del tipo *site* con *parasite*.

6.4.2.4 Cálculo del SRscore

El cálculo del *SRscore* se realiza a partir del nivel de parentesco, información adicional y conocimiento similar obtenidos, de acuerdo a lo explicado en el apartado 5.4.2.4. Para esto, ha sido necesario implementar la función de normalización *norm* (ver definición 5.21). Los resultados de esta implementación se mostrarán más adelante, en el apartado 6.4.5.

6.4.3 Evaluación de la popularidad

La implementación de la evaluación de la popularidad de cada ontología, consiste fundamentalmente en consultar varios recursos Web, para identificar el número de referencias que éstos contienen hacia la ontología que se pretende evaluar.

El número de referencias desde cada recurso a la ontología se almacenan en el repositorio de ontologías (campos *bioportalRefs*, *pubmedRefs*, *wikipediaRefs* y *twitterRefs* de la tabla *ontology*) en el momento en que la ontología se almacena en el repositorio, y se actualizan periódicamente según se desee. El conjunto de clases que realizan el acceso a los recursos Web para obtener el número de referencias están contenidas en el paquete *es.udc.imedir.java.collectiveknowledgeresources*, y se muestran en la figura 6.24.

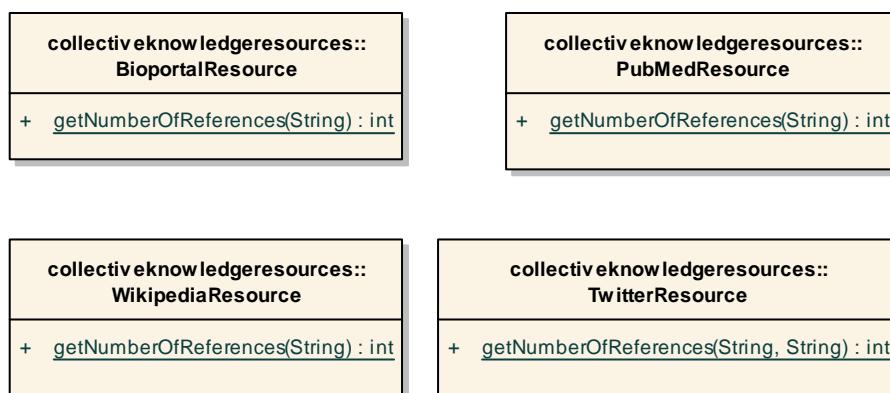


Figura 6.24. Conjunto de clases utilizadas para obtener las referencias desde cada recurso a una ontología.

A continuación, se explica cómo se ha obtenido esta información para cada uno de los cuatro recursos utilizados (i.e. BioPortal, PubMed, Wikipedia y Twitter).

En el caso de BioPortal, se trata de consultar la lista de ontologías que éste contiene para determinar si la ontología que se está evaluando figura en él o no. Debido a la ausencia de servicios para la recuperación de las ontologías de BioPortal, ha sido necesario parsear la sección de BioPortal en la que se listan las ontologías que contiene⁶² y comparar el nombre de la ontología en cuestión con los nombres de todas las ontologías en la lista. En el caso de que la ontología figure en la lista de BioPortal, el

⁶² <http://bioportal.bioontology.org/ontologies/>

número de referencias adopta un valor 1. En caso contrario, un 0. Para realizar el *parsing* del código HTML se ha utilizado la API Jericho⁶³, que permite “parsear” HTML de forma sencilla.

En el caso de PubMed, este recurso proporciona diversas utilidades o servicios que permiten acceder a la información que contiene a través de programación⁶⁴. Para este caso, se ha utilizado el servicio Web conocido como ESearch, que permite realizar diversas búsquedas sobre el material científico indexado en la BD de PubMed. Se ha buscado el nombre de la ontología, y almacenado el número de referencias obtenidas. A modo de ejemplo, en la figura 6.25 se muestra un caso de invocación del servicio para la ontología Gene Ontology. La salida obtenida se puede ver en la figura 6.26. Como indica el atributo *count*, PubMed contiene 3017 artículos que contienen el texto Gene Ontology. Éste será el número de referencias tenidas en cuenta hacia la ontología para este recurso.

```
http://www.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=%22gene%20ontology%22[All%20fields]
```

Figura 6.25. URL utilizada para buscar referencias a la ontología Gene Ontology en la BD de PubMed.

Para el acceso a Wikipedia, se utiliza la API proporcionada por MediaWiki⁶⁵. El acceso se realiza de forma similar al acceso a PubMed. En la figura 6.27 se muestra la URL que habría que utilizar para obtener el número de referencias a Gene Ontology. Los resultados obtenidos se pueden ver en la figura 6.28. El atributo *totalbits* indica que en Wikipedia existen 82 títulos de páginas o contenidos de páginas que contienen el texto Gene Ontology.

⁶³ <http://jericho.htmlparser.net/>

⁶⁴ <http://eutils.ncbi.nlm.nih.gov/>

⁶⁵ <http://en.wikipedia.org/w/api.php/>

```

<?xml version="1.0" ?>
<!DOCTYPE eSearchResult (View Source for full doctype...)>
http://www.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=%22gene%20ontology%22%5bAll%20fields%5d</eSearchResult>
<Count>3017</Count>
<RetMax>20</RetMax>
<RetStart>0</RetStart>
<IdList>
    <Id>21209420</Id>
    <...>
    <Id>21071420</Id>
</IdList>
<TranslationSet />
<TranslationStack>
<TermSet>
    <Term>"gene ontology"[All fields]</Term>
    <Field>All fields</Field>
    <Count>3017</Count>
    <Explode>Y</Explode>
</TermSet>
<OP>GROUP</OP>
</TranslationStack>
<QueryTranslation>"gene ontology"[All fields]</QueryTranslation>
</eSearchResult>

```

Figura 6.26. Resultado obtenido para la búsqueda de Gene Ontology en PubMed.

```

http://en.wikipedia.org/w/api.php?action=query&list=search&srsearch=%22gene%20ontology%22&srlimit=1&format=xml

```

Figura 6.27. URL utilizada para buscar referencias a la ontología Gene Ontology en Wikipedia.

```

<?xml version="1.0" ?>
<api>
    <query>
        <searchinfo totalhits="82" />
        <search>
            <...>
        </search>
    </query>
    <query-continue>
        <search sroffset="1" />
    </query-continue>
</api>

```

Figura 6.28. Resultado obtenido para la búsqueda de Gene Ontology en Wikipedia.

Finalmente, el acceso a Twitter se realiza mediante la API que proporciona el propio recurso⁶⁶, y que permite obtener el número de entradas (*o tweets*) en Twitter con el texto indicado. Continuando con el ejemplo anterior, la URL que habría que introducir para buscar el número de referencias a Gene Ontology sería la que se

⁶⁶ <http://apiwiki.twitter.com/w/page/22554679/Twitter-API-Documentation/>

muestra en la figura 6.29. El resultado obtenido se resume en la figura 6.30. La obtención del número de referencias consistiría en contar el número de entradas (`<entry>...</entry>`). Para el caso de Gene Ontology, el resultado es 11.

```
http://search.twitter.com/search.atom?q="gene ontology"&rpp=100
```

Figura 6.29. URL utilizada para buscar referencias a la ontología Gene Ontology en Twitter.

```
<?xml version="1.0" encoding="UTF-8"?>
<feed xmlns:google="http://base.google.com/ns/1.0" xml:lang="en-US"
      xmlns:openSearch="http://a9.com/-/spec/opensearch/1.1/"
      xmlns="http://www.w3.org/2005/Atom" xmlns:twitter="http://api.twitter.com/">
  <id>tag:search.twitter.com,2005:search/"gene ontology"</id>
  <entry>
    <...>
  </entry>
  <...>
  <entry>
    <...>
  </entry>
</feed>
```

Figura 6.30. Resultado obtenido para la búsqueda de "Gene Ontology" en Twitter.

6.4.4 Agregación de puntuaciones

A partir de los valores obtenidos tras evaluar cada ontología de acuerdo a los tres criterios propuestos (cobertura del contexto, riqueza semántica y popularidad), se obtiene la puntuación final para la ontología, agregando estos valores de acuerdo a lo explicado en el apartado 5.4.4. Esta agregación depende de los valores de un conjunto de pesos (tres pesos, uno para cada criterio), que ha sido necesario ajustar para poder ejecutar el prototipo de sistema de selección. El método seguido para ajustar estos pesos se explicará en el apartado 6.4.6.

6.4.5 Normalización de valores

En este apartado, se explica cómo se ha implementado la función de normalización *norm*, que pretende trasladar un valor en el intervalo abierto $[0, +\infty)$ al intervalo $[0, 1]$, teniendo en cuenta la distribución de una muestra de referencia. Para llevar a cabo la normalización de cada parámetro (e.g. nº de parientes de un concepto), se han realizado los siguientes pasos:

1. Selección de una muestra de referencia (e.g. número de parientes para cada concepto del repositorio de ontologías) y obtención de sus valores. La muestra se ha trasladado al software MATLAB⁶⁷, en forma de un *array* de una dimensión, que ha facilitado la realización de los pasos siguientes.
2. Eliminación del 0 en la muestra. Esto se ha realizado usando el siguiente comando de MATLAB: `nombreArray(nombreArray == 0) = []`.
3. División de la muestra en intervalos de igual frecuencia. Tras calcular el número de intervalos más adecuado siguiendo la regla de Sturges (Daniel & Wayne, 2009), se ha utilizado la función `prctile` de MATLAB para obtener los puntos que dividen a la muestra en intervalos de igual frecuencia. Los intervalos solapados se han agrupado en un único intervalo.
4. Finalmente, a cada uno de los intervalos obtenidos se le asigna un valor en el intervalo $[0, 1]$. Al i -ésimo intervalo se le asigna el valor i/k , con $1 \leq i \leq k$, donde k es el número de intervalos en que se ha dividido la muestra.

La tabla 6.5 contiene los datos de normalización correspondientes a los parámetros de la aproximación que es necesario normalizar. Estos parámetros son los siguientes:

Para el cálculo del *SRscore* (riqueza semántica):

- **Índice de parentesco un concepto en una ontología.** Se toma como muestra de referencia el número de parientes de cada concepto en el repositorio de ontologías.
- **Índice de información adicional un concepto en una ontología.** Se toma como muestra de referencia el número de elementos de información adicional de cada concepto en el repositorio.
- **Índice de conocimiento similar de un concepto en una ontología.** Se crea una muestra de referencia, obtenida calculando el número de conceptos similares para cada concepto del repositorio, en una ontología del repositorio elegida de forma aleatoria para cada concepto.

⁶⁷ <http://www.mathworks.com/products/matlab/>

Para el cálculo del *Pscore* (popularidad):

- **Número de referencias a una ontología desde BioPortal, PubMed, Wikipedia o Twitter.** Como muestra se toma el número de referencias a la ontología desde BioPortal, PubMed, Wikipedia o Twitter, para todas las ontologías del repositorio.

Tabla 6.5. Datos del proceso de normalización.

Parámetro	Muestra		Intervalos de frecuencia	
	Tamaño	Rango	Nº	Puntos de división
Índice de parentesco	1.860.911	[0, 13264]	18	1, 2, 3, 4, 5, 7, 8, 10, 13, 18, 24, 35, 55, 94, 195, 556, 2782
Índice de información adicional	1.860.911	[0, 534]	9	1, 2, 3, 4, 6, 8, 11, 17
Índice de conocimiento similar	1.860.911	[0, 86]	5	1, 2, 4, 6
Nº referencias desde BioPortal	200	[0, 1]	2	1
Nº referencias desde PubMed	200	[0, 24590]	6	2, 4, 13, 29, 60
Nº referencias desde Wikipedia	200	[0, 774]	6	1, 2, 4, 7, 23
Nº referencias desde Twitter	200	[0, 16]	5	1, 3, 4, 10

Ejemplo 6.1. Teniendo en cuenta los datos de la tabla 6.5, supóngase que se desea normalizar al rango [0, 1] un valor de 31 para el parámetro “nº de referencias desde PubMed”. A partir de los puntos de división se calculan los intervalos, y se asigna a cada uno de ellos un valor discreto en el intervalo [0, 1] calculado como i/k , con $1 \leq i \leq k$, donde k es el número de intervalos en que se ha dividido la muestra. Al 0 se le asigna siempre el valor 0. Los intervalos resultantes y sus valores correspondientes en el intervalo [0, 1] se muestran en la tabla 6.6.

Tabla 6.6. Intervalos resultantes y valores correspondientes en el intervalo [0, 1].

Intervalo	Valor
(0, 2)	$1/6 = 0,167$
[2, 4)	$2/6 = 0,333$
[4, 13)	$3/6 = 0,500$
[13, 29)	$4/6 = 0,667$
[29, 60)	$5/6 = 0,833$
[60, $+\infty$)	$6/6 = 1,000$

Observando la tabla, se puede ver que a 31 referencias en PubMed le correspondería un valor de 0,833 en el intervalo [0, 1], de acuerdo a la muestra tomada como base.

6.4.6 Ajuste de pesos

La aproximación utiliza varias variables que actúan como “pesos”, y cuya utilidad es la de dar mayor o menor importancia a los diferentes criterios utilizados durante el proceso de evaluación (ver apartado 5.4.6).

Aunque el prototipo construido permite ajustar los pesos utilizados según las características del contexto o problema que se pretende resolver, la evaluación del prototipo ha requerido determinar un conjunto de valores para estos pesos, a utilizar en un contexto de descripción semántica general, sin requerimientos especiales. Así, el ajuste de los pesos se ha realizado en base a la opinión de 5 expertos en el ámbito de las ontologías biomédicas, que han transmitido su opinión acerca de los valores a utilizar, a través de un *Formulario de ajuste de pesos* (ver anexo II).

Tabla 6.7. Puntuaciones proporcionadas por expertos (valores medios) para los pesos que intervienen en el cálculo de la puntuación final, riqueza semántica y popularidad de una ontología.

Cálculo	Criterio	Peso
Puntuación final	Cobertura del contexto	0,55
	Riqueza semántica	0,26
	Popularidad	0,19
Riqueza semántica	Parentesco	0,33
	Conocimiento similar	0,34
	Información adicional	0,33
Popularidad	Wikipedia	0,26
	Twitter	0,16
	BioPortal	0,28
	PubMed	0,30

En la tabla 6.7 se muestran las puntuaciones medias proporcionadas por el conjunto de expertos para los pesos de los que depende la aproximación. En cuanto a los pesos para el cálculo de la puntuación final, resalta el hecho de que los expertos consideran que el criterio de cobertura del contexto es, con diferencia, el más relevante.

Otro aspecto destacable es que el recurso Twitter es el que se ha considerado menos relevante para medir la popularidad de ontologías biomédicas.

6.5 Combinación y ordenación de ontologías

Como se ha explicado en el apartado 5.5, la aproximación propuesta dispone de dos tipos de salida: (1) Salida simple, que consiste en un ranking en el que cada elemento es una ontología, y que se realiza en base a la puntuación final obtenida para cada ontología durante el proceso de evaluación. (2) Salida combinada, en la que cada elemento del ranking puede estar formado una ontología o por un conjunto de ontologías.

A nivel de implementación, la salida del prototipo se expresa como una lista de elementos del tipo *OutputElement* (ver figura 6.31), clase perteneciente al paquete `es.udc.imedir.java.evaluator.facade`. Esta clase almacena la siguiente información sobre cada elemento de salida:

- **ontologyIds.** Identificadores de BD para la ontología u ontologías que constituyen el elemento de salida.
- **ontologyNames.** Nombres de las ontologías.
- **ccScore.** Puntuación obtenida para la ontología u ontologías tras el proceso de evaluación de cobertura del contexto.
- **srScore.** Puntuación obtenida para la ontología u ontologías tras el proceso de evaluación de riqueza semántica.
- **pScore.** Puntuación obtenida para la ontología u ontologías tras el proceso de evaluación de popularidad.
- **aggregatedScore.** Puntuación final para el elemento de salida, obtenida tras agregar las tres puntuaciones anteriores.
- **termsMatched.** Términos de entrada cubiertos por la ontología u ontologías.

Se trata de una lista de cadenas de texto, cada una de las cuales sigue el formato “term1 (term2)”, donde term1 es un término de entrada y term2 es el nombre del concepto que cubre el término de entrada (e.g. leukocyte(white cell)). El tamaño de esta lista es igual al número de términos de entrada (tras la

normalización, corrección y eliminación de sinónimos). Para cada término de entrada sin una correspondencia en las ontologías, la lista almacenará un *null*. De esta manera, a partir de los términos de entrada, esta lista permite conocer también los términos no cubiertos.

- **termsSources.** Lista, de igual tamaño que la anterior, que en cada posición almacena el identificador de la ontología que cubre el término de entrada.

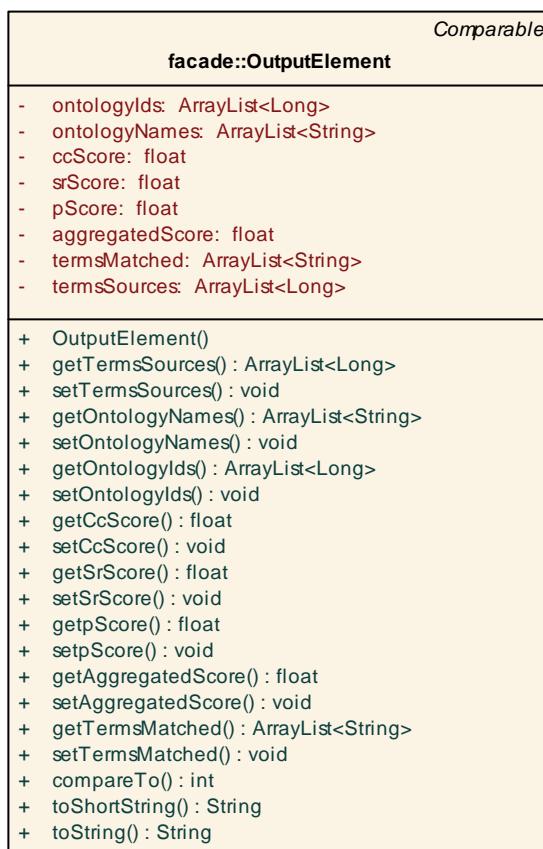


Figura 6.31. Clase *OutputElement*. Para simplificar la figura, se han ocultado los parámetros de los métodos.

En la figura 6.31 se puede observar también que la clase implementa la interfaz *Comparable*. El propósito de esto es definir un método para la comparación de dos elementos del tipo *OutputElement* cualesquiera, que permita ordenarlos y obtener el ranking final. Éste es el método *compareTo*, cuya implementación en lenguaje Java se puede observar en la figura 6.32. Se puede ver que la ordenación se realiza en base a la puntuación final (*AggregatedScore*) del elemento de salida. Elementos de salida (i.e. ontologías o conjuntos de ontologías) con mayor puntuación final, alcanzarán

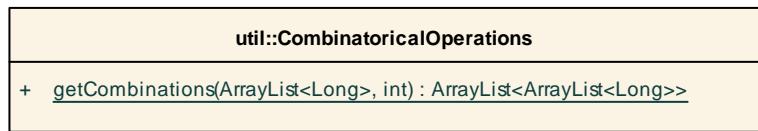
puntuaciones más altas en el ranking. En caso de empate, tienen prioridad las combinaciones de ontologías que mayor cobertura del contexto proporcionan. Y en caso de un nuevo empate, tienen prioridad las combinaciones en las que intervienen un menor número de ontologías. Con este algoritmo de comparación, se pretende obtener las mejores puntuaciones minimizando en la medida de lo posible el número de ontologías de cada resultado.

```
public int compareTo(Object o1) {
    if (this.getAggregatedScore() == ((OutputElement) o1).getAggregatedScore()) {
        if ((this.getCcScore() == ((OutputElement) o1).getCcScore())) {
            if ((this.getOntologyIds().size() ==
                ((OutputElement) o1).getOntologyIds().size()))
                return 0;
            else
                if ((this.getOntologyIds().size() >
                    ((OutputElement) o1).getOntologyIds().size()))
                    return 1;
                else
                    return -1;
        }
        else
            if ((this.getCcScore() < ((OutputElement) o1).getCcScore()))
                return 1;
            else
                return -1;
    }
    else
        if (this.getAggregatedScore() < ((OutputElement) o1).getAggregatedScore())
            return 1;
        else
            return -1;
}
```

Figura 6.32. Código en lenguaje Java del método *compareTo*, que permite comparar dos instancias del tipo *OutputElement*.

Finalmente, también resulta de interés explicar en este apartado cómo se han obtenido las combinaciones de ontologías. Esto se ha conseguido a través de la clase *CombinatorialOperations* del paquete `es.udc.imedir.java.bioss.math.util` (ver figura 6.33), que permite calcular todas las combinaciones sin repetición, de un determinado tamaño máximo, para un conjunto de identificadores de ontologías. Para ello, esta clase hace uso de la librería de clases Orbital⁶⁸, que proporciona diversas operaciones lógicas y matemáticas.

⁶⁸ <http://symbolaris.com/orbital/>

Figura 6.33. Clase *CombinatorialOperations*.

6.6 Interfaces proporcionadas

El prototipo proporciona dos interfaces de acceso: (1) Una interfaz de servicio Web, que hace posible su invocación de forma automática. (2) Una interfaz gráfica Web, a través de la que un usuario puede realizar el proceso de selección de ontologías. A continuación, se explica en qué consisten estas interfaces.

6.6.1 Servicio Web

Esta interfaz hace posible la invocación del sistema de selección de ontologías por parte de otros procesos, de tal manera que se permite su utilización desde programación a otros usuarios, o a agentes software que deseen acceder al servicio Web automáticamente en un contexto de reutilización automática de ontologías. El servicio proporciona las siguientes operaciones:

- **getOntologies.** Permite obtener toda la información de las ontologías del repositorio.
- **getExpandedTermsSets.** Realiza la expansión semántica de los términos de entrada.
- **selectOntologies.** A partir de los términos expandidos y un conjunto de parámetros de entrada (valores de pesos, tipo de salida, umbral mínimo de cobertura del contexto, etc.), realiza el proceso de selección, proporcionando como salida una lista ordenada de resultados (ontologías y sus puntuaciones).

Este servicio Web se ha implementado como un servicio SOAP⁶⁹ utilizando la versión Java del popular motor de servicios Web Apache Axis2⁷⁰. Para su despliegue,

⁶⁹ <http://www.w3.org/TR/soap/>

⁷⁰ <http://axis.apache.org/axis2/java/core/>

se ha utilizado el contenedor Apache Tomcat⁷¹. En el anexo III de este documento se puede ver el fichero WSDL (*Web Services Description Language*) del servicio, que proporciona una descripción detallada de la interfaz pública del servicio Web.

6.6.2 Interfaz gráfica

Además de la interfaz de servicio Web, el prototipo proporciona una interfaz gráfica Web, que se encuentra disponible públicamente en la dirección <http://bioss.ontologyselection.com>. Esta interfaz, desarrollada utilizando la tecnología .NET de Microsoft⁷², delega en el servicio Web mencionado anteriormente y permite ejecutar el proceso de selección de ontologías de manera intuitiva y con una gran flexibilidad. Como se puede ver en la figura 6.34, la interfaz dispone de un campo de texto en el que el usuario debe introducir las palabras clave que representan el contexto que se desea describir semánticamente, y varias opciones de entrada y de salida.

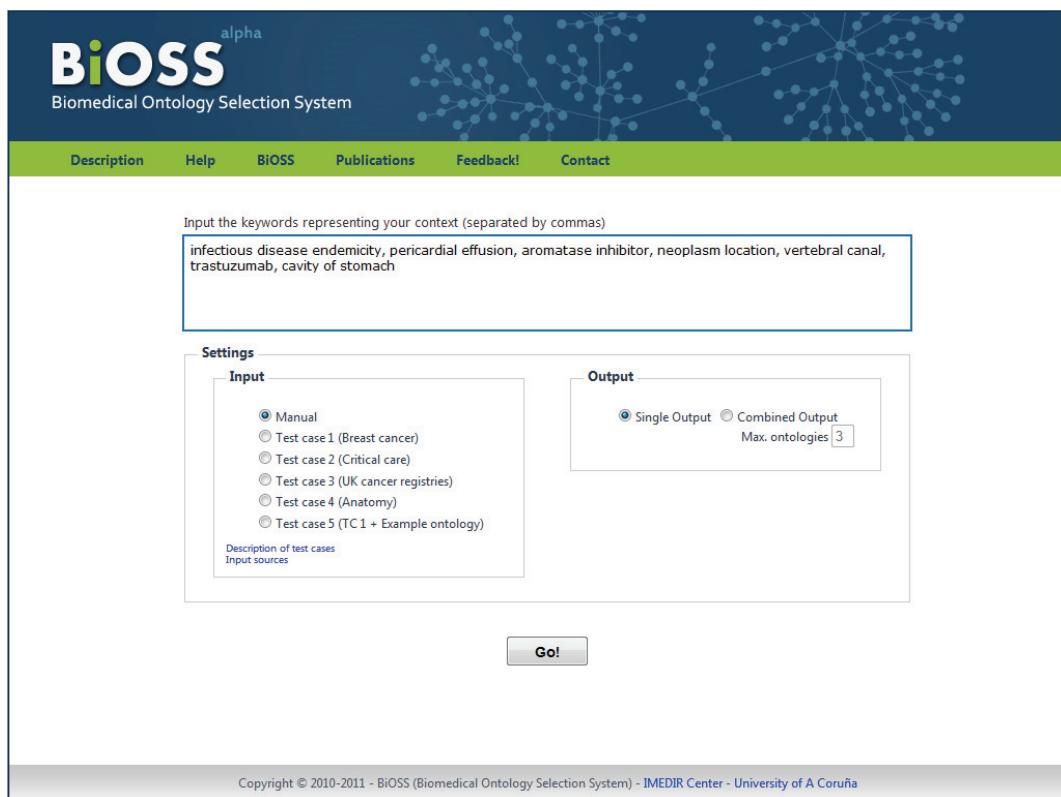


Figura 6.34. Interfaz gráfica del prototipo de sistema para la selección de ontologías biomédicas.

⁷¹ <http://tomcat.apache.org/>

⁷² <http://www.microsoft.com/net/>

Como opciones de entrada, se permite una entrada manual, a libertad del usuario, o bien alguno de los casos de estudio que se han utilizado para evaluar la aproximación (la evaluación se tratará en el capítulo siguiente). En cuanto a la salida, el usuario puede decidir entre la salida simple (cada resultado es una ontología) o la salida combinada (cada resultado puede ser una ontología o bien una combinación de dos o más ontologías). Para esta última, se puede modificar el número máximo de ontologías a combinar en cada resultado.

En la figura 6.35 se muestra un fragmento de la salida simple, para cinco términos de ejemplo: *white cell*, *chemotherapy*, *DNA*, *apoptosis* y *cavity of stomach*. Se puede ver que cada posición del ranking está ocupada por una única ontología, para la cual se muestra su nombre, la puntuación final obtenida (*Final score*) y los valores de *CCscore (Coverage)*, *SRscore (Richness)* y *Pscore (Popularity)* que han dado lugar a dicha puntuación (nótese que el *CCscore* se expresa en forma de porcentaje). También se muestran los términos iniciales cubiertos y no cubiertos por la ontología. Para los términos cubiertos, y en el caso de que el nombre del concepto de la ontología que proporciona la cobertura sea diferente del término cubierto, se indica entre paréntesis el nombre de dicho concepto.

RESULTS							
Position	Sources	Final score	Covered	Not covered	Coverage	Richness	Popularity
1	Medical Subject Headings	0.759	chemotherapy(pharmacother) white cell(white blood corpuscles) apoptosis(programmed cell death, type i)	cavity of stomach DNA	60.0%	0.918	1.000
2	NCI Thesaurus	0.737	chemotherapy white cell(wbc) apoptosis(pcd)	cavity of stomach DNA	60.0%	0.934	0.863
3	Logical Observation Identifier Names and Codes	0.709	chemotherapy white cell(leukocytes) DNA	cavity of stomach apoptosis	60.0%	0.951	0.697
4	Alcohol and Other Drug Thesaurus	0.648	chemotherapy(drug therapy) white cell(leukocytes) apoptosis DNA	cavity of stomach	80.0%	0.801	0.000
5	Gene Ontology	0.635	apoptosis(type i programmed cell death) DNA(dna location)	chemotherapy white cell cavity of stomach	40.0%	0.867	1.000
6	CRISP Thesaurus	0.603	chemotherapy(drug therapy) white cell(wbc (white blood cell)) apoptosis(pcd)	cavity of stomach DNA	60.0%	0.918	0.180
7	Foundational Model of Anatomy	0.523	white cell(leukocyte) cavity of stomach	chemotherapy apoptosis DNA	40.0%	0.834	0.453
8	Experimental Factor Ontology	0.520	white cell(leukocyte) DNA	chemotherapy cavity of stomach apoptosis	40.0%	0.951	0.280
9	Cancer Research and Management ACGT Master Ontology	0.470	chemotherapy white cell(leucocyte) DNA	cavity of stomach apoptosis	60.0%	0.302	0.323
10	Malaria Ontology	0.468	chemotherapy white cell(leukocyte)	cavity of stomach apoptosis DNA	40.0%	0.719	0.323
11	Galen Ontology	0.449	chemotherapy white cell(leucocyte) apoptosis	cavity of stomach DNA	60.0%	0.255	0.280
12	Cardiac Electrophysiology Ontology	0.446	white cell(leukocyte) cavity of stomach	chemotherapy apoptosis DNA	40.0%	0.664	0.280
13	HL7 Vocabulary Version 3.0	0.413	chemotherapy(drug therapy) white cell(leukocytes) apoptosis(pcd)	cavity of stomach DNA	60.0%	0.318	0.000
14	Online Mendelian Inheritance in Man	0.386	apoptosis(pcd)	chemotherapy white cell cavity of stomach DNA	20.0%	0.449	0.840

Figura 6.35. Ejemplo de salida simple.

La figura 6.36 muestra la salida combinada para el mismo conjunto de términos de entrada. En este caso, cada posición del ranking está ocupada por una o por varias ontologías (en este caso siempre 3, pero puede variar), que se combinan para proporcionar mejores valores de evaluación. En la salida simple, la ontología que ocupa la primera posición del ranking proporciona una cobertura del contexto de un 60%. La salida combinada muestra que es posible alcanzar una cobertura de un 100% mediante la combinación de varias ontologías, con valores de puntuación final superiores a 0,9. Obsérvese que, en la salida combinada, el nombre de cada ontología para cada posición del ranking se encuentra representado usando un color diferente, y que este color se utiliza para indicar cuál es la ontología que proporciona la cobertura para cada término cubierto. Por ejemplo, para la primera posición del ranking, se puede ver que uno de los términos cubiertos por la ontología Medical Subject Headings (en azul) es *white cell*, a través del concepto de la ontología de nombre *white blood corpuscles*.

RESULTS							
Position	Sources	Final score	Covered	Not covered	Coverage	Richness	Popularity
1	Medical Subject Headings Gene Ontology Foundational Model of Anatomy	0.951	chemotherapy(pharmacother) white cell(white blood corpuscles) cavity of stomach apoptosis(programmed cell death, type i) DNA(dna location)		100.0%	0.891	0.891
2	NCI Thesaurus Gene Ontology Foundational Model of Anatomy	0.938	chemotherapy white cell(wbc) cavity of stomach apoptosis(pcd) DNA(dna location)		100.0%	0.901	0.809
3	Medical Subject Headings Gene Ontology Cardiac Electrophysiology Ontology	0.935	chemotherapy(pharmacother) white cell(white blood corpuscles) cavity of stomach apoptosis(programmed cell death, type i) DNA(dna location)		100.0%	0.857	0.856
4	Medical Subject Headings Gene Ontology University of Washington Digital Anatomist	0.926	chemotherapy(pharmacother) white cell(white blood corpuscles) cavity of stomach(lumen of stomach) apoptosis(programmed cell death, type i) DNA(dna location)		100.0%	0.849	0.820
5	Medical Subject Headings Logical Observation Identifier Names and Codes Foundational Model of Anatomy	0.924	chemotherapy white cell(leukocytes) cavity of stomach apoptosis(programmed cell death, type i) DNA		100.0%	0.921	0.709
6	NCI Thesaurus Gene Ontology Cardiac Electrophysiology Ontology	0.922	chemotherapy white cell(wbc) cavity of stomach apoptosis(pcd) DNA(dna location)		100.0%	0.867	0.774
7	Logical Observation Identifier Names and Codes Gene Ontology Foundational Model of Anatomy	0.921	chemotherapy white cell(leukocytes) cavity of stomach apoptosis(type i programmed cell death) DNA		100.0%	0.910	0.709
8	NCI Thesaurus Logical Observation Identifier Names and Codes Foundational Model of Anatomy	0.920	chemotherapy white cell(leukocytes) cavity of stomach apoptosis(pcd) DNA		100.0%	0.924	0.681
9	Medical Subject Headings Foundational Model of Anatomy FlyBase Controlled Vocabulary	0.919	chemotherapy(pharmacother) white cell(white blood corpuscles) cavity of stomach apoptosis(programmed cell death, type i) DNA		100.0%	0.874	0.747
10	Medical Subject Headings Foundational Model of Anatomy Subcellular Anatomy Ontology	0.918	chemotherapy(pharmacother) white cell(white blood corpuscles) cavity of stomach apoptosis(programmed cell death, type i) DNA		100.0%	0.864	0.755

Figura 6.36. Ejemplo de salida combinada.

7 Evaluación

Para evaluar la aproximación propuesta en esta tesis, se ha realizado un experimento en el dominio biomédico, utilizando el prototipo de sistema de selección de ontologías que implementa dicha aproximación⁷³, descrito en el capítulo 6. Este experimento se describe a continuación.

7.1 Descripción del experimento

El experimento ha consistido en realizar un estudio de casos en el que, con la ayuda de varios expertos, se han evaluado los resultados proporcionados por el prototipo en cinco escenarios (casos de estudio) representativos del ámbito de la selección de ontologías biomédicas. Los casos de estudio han sido los siguientes:

- **Caso de estudio 1 (cáncer de mama):** Un equipo de investigación de la Universidad de Stanford (EE.UU.) está construyendo un sistema que abstrae información clínica de dos bases de datos médicas relacionadas con el cuidado y gestión del cáncer de mama. El objetivo de este trabajo es medir la calidad de la atención y el grado de aplicación de las guías clínicas, como se describe en la National Comprehensive Cancer Network⁷⁴. Para la construcción de esta aplicación, este conjunto de investigadores necesita reutilizar ontologías ya desarrolladas por otras organizaciones.
- **Caso de estudio 2 (cuidados críticos):** En el Centro IMEDIR de la Universidad de A Coruña, interesa disponer de una ontología con los

⁷³ Disponible en <http://bioss.ontologyselection.com/>

⁷⁴ <http://www.nccn.org/>

principales términos utilizados en el ámbito de las unidades de cuidados críticos (UCIs), durante la monitorización de pacientes que han sido intervenidos de cirugía cardíaca. Esta ontología se integrará en un sistema de apoyo a la toma de decisión en una de las unidades de cuidados críticos del hospital do Meixoeiro⁷⁵ de la ciudad de Vigo. Para desarrollar esta ontología, interesa reutilizar el conocimiento de ontologías ya existentes. Este trabajo se enmarca dentro de los proyectos de investigación FIS-PI061524 y FIS-PI10/02180, financiados por el Instituto de Salud Carlos III, y del proyecto XUNTA-2007/127, financiado por la Dirección General Promoción Científica y Tecnológica del Sistema Universitario de Galicia, de la Xunta de Galicia.

- **Caso de estudio 3 (registro de cáncer del Reino Unido):** El Reino Unido está cubierto por 11 registros sobre cáncer, coordinados por la Asociación de Registros de Cáncer del Reino Unido (United Kingdom Association of Cancer Registries, UKACR)⁷⁶. Estos registros, que almacenan información sobre cáncer poblacional desde hace más de 40 años, recopilan el mismo conjunto mínimo de datos, formado por 24 variables. En este caso de estudio se supone que se desea buscar las mejores ontologías para anotar semánticamente estas variables, pensando en un posible futuro caso de integración de información epidemiológica.
- **Caso de estudio 4 (anatomía):** Para este caso se supone que se desea encontrar las mejores ontologías para describir semánticamente un conjunto de términos del dominio de la anatomía, elegidos aleatoriamente.
- **Caso de estudio 5 (ontología de ejemplo):** Este último caso de estudio ha consistido en introducir en el repositorio de ontologías una ontología creada de forma individual, no publicada ni consensuada con el resto de la comunidad biomédica, que proporciona una cobertura de un contexto determinado (se ha utilizado como referencia el caso de estudio 1) superior al de todas las demás ontologías del repositorio para tan contexto. El objetivo es observar la capacidad de la aproximación para descartar esta ontología por tratarse de una ontología no popular y con baja riqueza semántica.

⁷⁵ <http://tinyurl.com/38k29wl>

⁷⁶ <http://www.ukacr.org/>

En la tabla 7.1 se muestran los términos de entrada para los casos de estudio. Se han usado conjuntos de términos sin errores y evitando la introducción de sinónimos.

Tabla 7.1. Términos de entrada utilizados en los casos de estudio (CE).

Caso	Términos de entrada	Nº términos
CE 1	ductal carcinoma in situ, adjuvant chemotherapy, axillary lymph node staging, mastectomy, tamoxifen, serotonin reuptake inhibitors, invasive breast cancer, hormone receptor positive breast cancer, ovarian ablation, premenopausal women, surgical management, biopsy of breast tumor, fine needle aspiration, sentinel lymph node, breast preservation, adjuvant radiation therapy, prechemotherapy, inflammatory breast cancer, ovarian failure, bone scan, lumpectomy, brain metastases, pericardial effusion, aromatase inhibitor, postmenopausal, palliative care, guidelines, stage iv breast cancer disease, trastuzumab, breast mri examination	30
CE 2	vital signs, mean arterial pressure, cardiac output, pressure venous central, expert systems, decision, serum, quadruple, add, stop, double, ascend, descend, continuous, medical device, pump, monitoring device, pharmacologic substance, vasodilator, nitroprusside, nitroglycerin, adrenaline, inotropic agents, noradrenaline, dobutamine, accepted, lower limit value, upper limit value, value, name, priority, variation	32
CE 3	hospital, consultant, patient unit number, nhs number, forenames, surname, name at birth, address at time of diagnosis, postal code, sex, ethnic origin, date of birth, neoplasm location, morphology, laterality, stage, tumor grading, basis of diagnosis, date of diagnosis, treatment indicators, vital status, date of death, cause and place of death, post mortem	24
CE 4	dimorphism, biliary system, distal, dorsal, ectoderm, electrolyte, sural nerve, synapse, calcaneal tendon, endocardium, endoderm, epithelium, gonadotropins, heterophilic, homeothermic, squamous, stratified epithelium, set of meninges, paraganglion, radial nerve, dopamine receptor, syncytium, telogen, skull, cavity of stomach, surface of bone of foot, cranial cavity, spinal canal, thoracic cavity, gingiva, abdominopelvic cavity, spectrin, pericardial cavity, gallbladder, spleen, kidneys, set of muscles of abdomen, basilar artery, hepatic vein, uterine tubes, sagittal plane, ventral, vertebral canal, agger nasi, allantois, alveus, anconeus, annulus, hamstrings, quadrigeminus, macula, mammilla, masseter, mediastinum, mesencephalon, modiolus, mylohyoid, pampiniform, paraesthesia, parotid, pedicle, pennate, periosteum, platysma, proprioceptive, psoas, arch of aorta, axon, lysosome, mechanoreceptor, tibial tuberosity, perimysium, musculature of head, rectus capitis anterior, popliteus, midcarpal joint, obturator canal, long bone, skin of elbow, umbilical vein, supination	80
CE 5	Como en el CE 1	

El sistema se ha utilizado para obtener las ontologías más adecuadas a cada caso de estudio. Los resultados obtenidos han sido analizados y comentados por el autor. Además, se ha contado con la opinión de varios expertos familiarizados con las principales ontologías biomédicas y habituados a utilizar las herramientas habituales para examinarlas. Estos expertos han evaluado, por una parte, los resultados obtenidos por el sistema para cada caso de estudio y, por otra, la utilidad del sistema de forma general. Para ello, han utilizado un formulario de evaluación *online*, que se puede ver en

el anexo II de este documento (ver *Formulario de evaluación por parte de expertos*). Este cuestionario tiene la siguiente estructura:

- **Sección 1. Información personal.** Preguntas en las que se recopila información personal sobre el experto (nombre y apellidos, institución, correo electrónico y país).
- **Sección 2. Casos de estudio.** Preguntas en relación a cada uno de los casos de estudio. Se pretende que el experto indique las ontologías que él considera más adecuadas para cada caso de estudio y luego opine sobre la salida obtenida por el sistema.
- **Sección 3. Preguntas generales.** Varias preguntas cuya finalidad es obtener la opinión del experto sobre la utilidad general del sistema, así como sus principales ventajas e inconvenientes.

Para todos los casos de estudio, la evaluación de los expertos se ha limitado a los 10 primeros resultados proporcionados por el sistema, y se ha tenido en cuenta tanto la salida simple como combinada (excepto en el CE 5, en el que únicamente se ha considerado la salida simple, dado que la salida combinada no proporciona información adicional respecto a lo que se pretende estudiar). En el caso de la salida de combinaciones de ontologías, se ha establecido en 5 el número máximo de ontologías que pueden formar parte de cada combinación. También se ha fijado un valor de 0,1 (un 10% de los términos iniciales) como mínimo de cobertura permitida para que una ontología participe en la evaluación.

A continuación, se presentan y analizan los resultados proporcionados por el sistema para cada caso de estudio. Posteriormente, se exponen los resultados de la evaluación realizada por los expertos.

7.2 Resultados proporcionados por el prototipo

En este apartado se presentan los resultados de ejecutar cada caso de estudio usando el prototipo. Nótese que al comentar los resultados y para facilitar la comprensión, los valores de cobertura de contexto se expresarán en forma de porcentajes. Por ejemplo,

se dirá que una ontología cubre un 70% de los términos iniciales, cuando su valor de cobertura del contexto es de 0,7.

7.2.1 Caso de estudio 1 (cáncer de mama)

En este caso de estudio se puede observar (ver tabla 7.2) que únicamente existen 8 ontologías del repositorio que proporcionan una cobertura del contexto superior al límite mínimo establecido (0,1, es decir, 10%). De estas 8 ontologías 2 son específicas del dominio del cáncer (NCI Thesaurus y Cancer Research Management ACGT Master Ontology), y 1 de ellas se centra en las células de mama (Breast Tissue Cell Lines Ontology). El NCI Thesaurus es la ontología que proporciona los mejores resultados, con una cobertura de un 70% de los términos iniciales y una puntuación final de 0,779.

Tabla 7.2. Resultados del prototipo para el caso de estudio 1 con salida simple.

Posición	Ontología	CCscore	SRscore	Pscore	Score final
1	NCI Thesaurus	0,700	0,884	0,863	0,779
2	Medical Subject Headings	0,433	0,868	1,000	0,654
3	Physician Data Query	0,200	0,768	0,703	0,443
4	MedlinePlus Health Topics	0,167	0,407	0,483	0,289
5	CRISP Thesaurus	0,300	0,322	0,180	0,283
6	Cancer Research and Management ACGT Master Ontology	0,200	0,173	0,323	0,216
7	Breast Tissue Cell Lines Ontology	0,133	0,239	0,280	0,189
8	Cell Line Ontology (MCCL)	0,133	0,239	0,280	0,189
9	-	-	-	-	-
10	-	-	-	-	-

En cuanto a la salida combinada (ver tabla 7.3), el uso conjunto del NCI Thesaurus y Medical Subject Headings permite alcanzar una cobertura del contexto de un 73%, con unos valores de riqueza semántica y popularidad muy elevados (0,88 y 0,87 respectivamente). Se puede observar que, aunque en la salida simple el NCI Thesaurus es la ontología que proporcionaba, con diferencia, la mayor cobertura (70% respecto al 43% de la segunda ontología en el ranking), existen combinaciones de ontologías que alcanzan valores de cobertura superiores al 50% prescindiendo del NCI Thesaurus (resultados 5 al 10) de la salida combinada. Estos resultados permiten intuir la potencia

que proporciona la salida combinada respecto a la salida simple, y la dificultad de llegar a estos resultados sin disponer de una herramienta como la que aquí se propone.

Tabla 7.3. Resultados del prototipo para el caso de estudio 1 con salida combinada.

Posición	Ontologías	CCscore	SRscore	Pscore	Score final
1	- NCI Thesaurus - Medical Subject Headings	0,733	0,883	0,870	0,798
2	- NCI Thesaurus - MedlinePlus Health Topics	0,733	0,862	0,846	0,788
3	- NCI Thesaurus - CRISP Thesaurus	0,733	0,858	0,832	0,785
4	- NCI Thesaurus	0,700	0,884	0,863	0,779
5	- Medical Subject Headings - Physician Data Query - CRISP Thesaurus - Breast Tissue Cell Lines Ontology	0,600	0,786	0,865	0,699
6	- Medical Subject Headings - Physician Data Query - CRISP Thesaurus - Cell Line Ontology (MCCL)	0,600	0,786	0,865	0,699
7	- Medical Subject Headings - Physician Data Query - CRISP Thesaurus	0,567	0,818	0,899	0,695
8	- Medical Subject Headings - Physician Data Query - Breast Tissue Cell Lines Ontology	0,567	0,813	0,905	0,695
9	- Medical Subject Headings - Physician Data Query - Cell Line Ontology (MCCL)	0,567	0,813	0,905	0,695
10	- Medical Subject Headings - Physician Data Query	0,533	0,849	0,944	0,694

7.2.2 Caso de estudio 2 (cuidados críticos)

En la tabla 7.4 se presentan las ontologías seleccionadas para este caso de estudio usando el modo de salida simple. Se puede observar que, al igual que en el anterior caso de estudio, el NCI Thesaurus es la ontología que encabeza la lista de resultados, con una cobertura del contexto de un 75%, elevados valores de riqueza semántica y popularidad (0,730 y 0,863 respectivamente) y un score final de 0,766. La siguiente ontología en el ranking es Medical Subject Headings. Existen otras ontologías que proporcionan una mayor cobertura del contexto que ella (e.g. Alcohol and Other Drug Thesaurus), pero Medical Subject Headings es una ontología con una gran relevancia

en la comunidad biomédica, y muy rica semánticamente. Estos valores la sitúan en la segunda posición.

Como ya se ha explicado, este caso de estudio se enmarca en el contexto de las unidades de cuidados críticos para la monitorización de pacientes que han sido intervenidos de cirugía cardíaca. La existencia de diversos términos de entrada relacionados con tratamientos cardíacos también motiva la aparición en la salida de una ontología dedicada a la electrofisiología cardíaca (Cardiac Electrophysiology Ontology).

Tabla 7.4. Resultados del prototipo para el caso de estudio 2 con salida simple.

Posición	Ontología	CCscore	SRscore	Pscore	Score final
1	NCI Thesaurus	0,750	0,730	0,863	0,766
2	Medical Subject Headings	0,344	0,868	1,000	0,605
3	Logical Observation Identifier Names and Codes	0,344	0,835	0,697	0,539
4	CRISP Thesaurus	0,375	0,800	0,180	0,448
5	Alcohol and Other Drug Thesaurus	0,406	0,676	0,000	0,399
6	Chemical Entities of Biological Interest Ontology	0,188	0,901	0,280	0,390
7	Cardiac Electrophysiology Ontology	0,219	0,730	0,280	0,363
8	International Classification for Nursing Practice	0,188	0,545	0,410	0,323
9	Galen Ontology	0,313	0,304	0,280	0,304
10	HL7 Vocabulary Version 2.5	0,250	0,611	0,000	0,296

En cuanto a la salida combinada, ésta se muestra en la tabla 7.5, y permite ver que existen 9 combinaciones de ontologías que mejoran notablemente los resultados de la salida simple, proporcionando una cobertura del contexto de un 87%. Todas ellas son combinaciones del NCI Thesaurus con otras ontologías.

7.2.3 Caso de estudio 3 (registro de cáncer)

En este caso de estudio, de nuevo el NCI Thesaurus es la ontología que lidera la lista de ontologías seleccionadas (ver tabla 7.6), lo cual tiene sentido considerando que los términos de entrada son las variables de un registro de cáncer.

Tabla 7.5. Resultados del prototipo para el caso de estudio 2 con salida combinada.

Posición	Ontologías	CCscore	SRscore	Pscore	Score final
1	- NCI Thesaurus - Medical Subject Headings	0,875	0,784	0,917	0,859
2	- NCI Thesaurus - Medical Subject Headings - Logical Observation Identifier Names and Codes	0,875	0,799	0,893	0,859
3	- NCI Thesaurus - Medical Subject Headings - Logical Observation Identifier Names and Codes - CRISP Thesaurus	0,875	0,802	0,869	0,855
4	- NCI Thesaurus - Medical Subject Headings - CRISP Thesaurus	0,875	0,789	0,868	0,851
5	- NCI Thesaurus - Medical Subject Headings - Chemical Entities of Biological Interest Ontology	0,875	0,796	0,768	0,834
6	- NCI Thesaurus - Medical Subject Headings - Logical Observation Identifier Names and Codes - Chemical Entities of Biological Interest -Ontology	0,875	0,811	0,744	0,833
7	- NCI Thesaurus - Medical Subject Headings - Logical Observation Identifier Names and Codes - CRISP Thesaurus - Chemical Entities of Biological Interest Ontology	0,875	0,813	0,719	0,829
8	- NCI Thesaurus - Medical Subject Headings - CRISP Thesaurus - Chemical Entities of Biological Interest Ontology	0,875	0,801	0,719	0,826
9	- NCI Thesaurus - Logical Observation Identifier Names and Codes - Alcohol and Other Drug Thesaurus - Library of Congress Subject Headings	0,875	0,744	0,749	0,817
10	- NCI Thesaurus - Logical Observation Identifier Names and Codes - Alcohol and Other Drug Thesaurus	0,844	0,771	0,763	0,809

En este caso, la cobertura del contexto del NCI Thesaurus desciende al 58%, debido a la ausencia de ontologías en el repositorio que contengan conceptos para representar términos complejos como *Address at time of diagnosis*, *Cause and place of death*,

etc. Estos términos podrían descomponerse en términos más simples previamente a realizar la selección. Sin embargo, esto queda fuera del alcance del presente trabajo.

Tabla 7.6. Resultados del prototipo para el caso de estudio 3 con salida simple.

Posición	Ontología	CCscore	SRscore	Pscore	Score final
1	NCI Thesaurus	0,583	0,595	0,863	0,639
2	Medical Subject Headings	0,333	0,850	1,000	0,594
3	Logical Observation Identifier Names and Codes	0,417	0,669	0,697	0,535
4	CRISP Thesaurus	0,250	0,801	0,180	0,380
5	Experimental Factor Ontology	0,167	0,787	0,280	0,349
6	Neural ElectroMagnetic Ontologies	0,167	0,705	0,330	0,338
7	Library of Congress Subject Headings	0,292	0,330	0,360	0,315
8	HL7 Vocabulary Version 2.5	0,292	0,570	0,000	0,309
9	Alcohol and Other Drug Thesaurus	0,167	0,768	0,000	0,291
10	Cancer Research and Management ACGT Master Ontology	0,292	0,239	0,323	0,284

La salida combinada (ver tabla 7.7) proporciona varias formas de mejorar la cobertura del contexto respecto a la salida simple (desde un 58% hasta un 67%), con valores para el score final superiores a 0,7 y, por lo tanto, siempre mejores que los obtenidos durante la salida simple. Para esto se requiere la intervención de 3 a 5 ontologías diferentes. En la mayoría de estas combinaciones intervienen el NCI Thesaurus y Medical Subject Headings.

Tabla 7.7. Resultados del prototipo para el caso de estudio 3 con salida combinada.

Posición	Ontologías	CCscore	SRscore	Pscore	Score final
1	- NCI Thesaurus - Medical Subject Headings - HL7 Vocabulary Version 2.5	0,667	0,721	0,878	0,721
2	- NCI Thesaurus - Medical Subject Headings - Logical Observation Identifier Names and Codes - HL7 Vocabulary Version 2.5	0,667	0,749	0,815	0,716
3	- NCI Thesaurus - Medical Subject Headings - Neural ElectroMagnetic Ontologies - HL7 Vocabulary Version 2.5	0,667	0,735	0,811	0,712
4	- NCI Thesaurus - Medical Subject Headings	0,625	0,731	0,936	0,712

5	- NCI Thesaurus - Medical Subject Headings - HL7 Vocabulary Version 2.5 - HL7 Vocabulary Version 3.0	0,667	0,722	0,824	0,711
6	- NCI Thesaurus - Medical Subject Headings - Logical Observation Identifier Names and Codes - Neural ElectroMagnetic Ontologies - HL7 Vocabulary Version 2.5	0,667	0,753	0,769	0,709
7	- NCI Thesaurus - Medical Subject Headings - Logical Observation Identifier Names and Codes	0,625	0,761	0,870	0,707
8	- Medical Subject Headings - Logical Observation Identifier Names and Codes - HL7 Vocabulary Version 2.5 - HL7 Vocabulary Version 3.0	0,667	0,750	0,761	0,706
9	- Medical Subject Headings - Logical Observation Identifier Names and Codes - HL7 Vocabulary Version 2.5 - Cancer Research and Management ACGT Master Ontology	0,667	0,726	0,781	0,704
10	- NCI Thesaurus - Medical Subject Headings - Neural ElectroMagnetic Ontologies	0,625	0,745	0,865	0,702

7.2.4 Caso de estudio 4 (anatomía)

En este caso de estudio la entrada es un conjunto de términos del dominio de la anatomía. En los resultados obtenidos con la salida simple (ver tabla 7.8) se puede ver que la ontología con mayor puntuación final es el Foundational Model of Anatomy, que es una de las ontologías más relevantes en el dominio de la anatomía. Esta ontología proporciona una cobertura del contexto de un 82%, que la lleva a encabezar el ranking aún por encima de ontologías más populares y ricas, como el NCI Thesaurus o Medical Subject Headings. También es de destacar que en el “top 10” figuran otras dos ontologías con una muy buena cobertura del contexto, como son la Cardiac Electrophysiology Ontology (82%) y University of Washington Digital Anatomist (70%), ambas de temáticas directamente relacionadas con la anatomía.

Tabla 7.8. Resultados del prototipo para el caso de estudio 4 con salida simple.

Posición	Ontología	CCscore	SRscore	Pscore	Score final
1	Foundational Model of Anatomy	0,825	0,850	0,453	0,761
2	NCI Thesaurus	0,625	0,934	0,863	0,751
3	Medical Subject Headings	0,513	0,817	1,000	0,684
4	Cardiac Electrophysiology Ontology	0,825	0,562	0,280	0,653
5	University of Washington Digital Anatomist	0,700	0,883	0,100	0,634
6	Logical Observation Identifier Names and Codes	0,263	0,967	0,697	0,528
7	Galen Ontology	0,538	0,271	0,280	0,419
8	BRENDA tissue / enzyme source	0,313	0,746	0,280	0,419
9	CRISP Thesaurus	0,363	0,683	0,180	0,411
10	Experimental Factor Ontology	0,238	0,835	0,280	0,401

Por lo que respecta a la salida combinada (ver tabla 7.9), existen varios conjuntos de ontologías que proporcionan una gran cobertura del contexto, con valores de riqueza semántica y popularidad elevados. Así, 7 de los 10 grupos de ontologías en el ranking proporcionan una cobertura conjunta de un 90%, con una puntuación final de 0,86.

Tabla 7.9. Resultados del prototipo para el caso de estudio 4 con salida combinada.

Posición	Ontologías	CCscore	SRscore	Pscore	Score final
1	- Foundational Model of Anatomy - NCI Thesaurus - Cardiac Electrophysiology Ontology - BRENDA tissue / enzyme source - CRISP Thesaurus	0,900	0,901	0,729	0,868
2	- Foundational Model of Anatomy - NCI Thesaurus - Cardiac Electrophysiology Ontology - BRENDA tissue / enzyme source - Alcohol and Other Drug Thesaurus	0,900	0,901	0,727	0,867
3	- Foundational Model of Anatomy - NCI Thesaurus - Galen Ontology - BRENDA tissue / enzyme source - CRISP Thesaurus	0,900	0,897	0,729	0,867
4	- Foundational Model of Anatomy - NCI Thesaurus - Cardiac Electrophysiology Ontology - Galen Ontology - BRENDA tissue / enzyme source	0,900	0,895	0,731	0,867

5	- Foundational Model of Anatomy - NCI Thesaurus - Galen Ontology - BRENDA tissue / enzyme source - Alcohol and Other Drug Thesaurus	0,900	0,897	0,727	0,866
6	- Foundational Model of Anatomy - NCI Thesaurus - Cardiac Electrophysiology Ontology - Galen Ontology - CRISP Thesaurus	0,900	0,894	0,729	0,866
7	- Foundational Model of Anatomy - NCI Thesaurus - Cardiac Electrophysiology Ontology - Galen Ontology - Alcohol and Other Drug Thesaurus	0,900	0,894	0,727	0,866
8	- Foundational Model of Anatomy - NCI Thesaurus - BRENDA tissue / enzyme source - CRISP Thesaurus	0,888	0,905	0,736	0,863
9	- Foundational Model of Anatomy - NCI Thesaurus - Cardiac Electrophysiology - Ontology - BRENDA tissue / enzyme source	0,888	0,904	0,737	0,863
10	- Foundational Model of Anatomy - NCI Thesaurus - BRENDA tissue / enzyme source - Alcohol and Other Drug Thesaurus	0,888	0,905	0,733	0,863

7.2.5 Caso de estudio 5 (ontología de ejemplo)

Este caso ha consistido en construir una ontología de ejemplo, con todos los términos de entrada utilizados en uno de los casos de estudio anteriores (se ha utilizado el caso de estudio 1), de modo que la cobertura del contexto sea del 100%. El objetivo es comprobar si el sistema es capaz de detectar que no se trata de una ontología relevante en la comunidad biomédica. Esta ontología se ha creado usando el software de creación de ontologías Protégé. Está formada por 30 conceptos hermanos, todos ellos hijos del concepto general *Thing*. En la figura 7.1 se muestra una captura de Protégé en la que se puede ver la jerarquía de conceptos de la ontología. Es evidente que se trata de una ontología con un detalle prácticamente nulo de información para cada uno de sus conceptos, por lo que su riqueza semántica será baja. Además, como esta ontología

no se ha publicado ni compartido con el resto de la comunidad biomédica para consensuar su conocimiento, el sistema debería asignarle una popularidad muy baja, o nula.

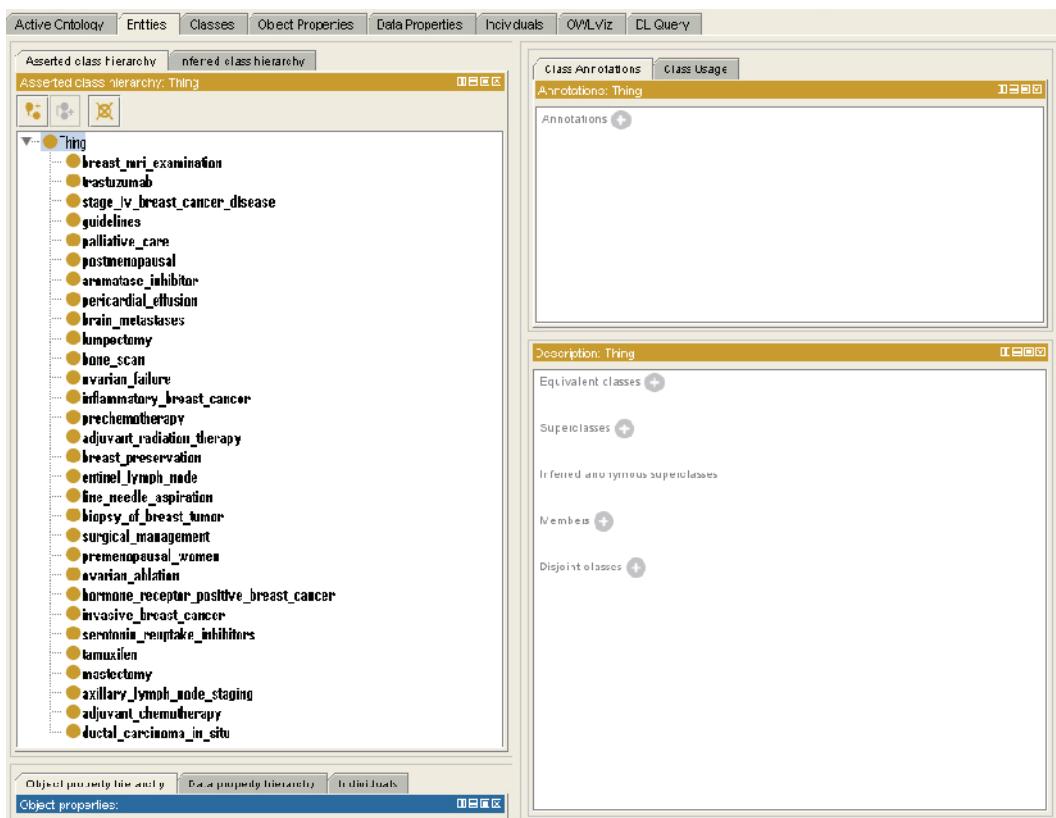


Figura 7.1. Captura de Protégé en la que se puede ver el árbol de conceptos de la ontología de ejemplo, con todos los conceptos del caso de estudio 1.

Observando la salida obtenida (ver tabla 7.10), se puede ver que aunque la ontología de ejemplo proporciona una cobertura de un 100%, pues cubre todos los términos de entrada, ésta ha quedado relegada a la tercera posición del ranking debido a su baja riqueza semántica y nula popularidad. Está por debajo del NCI Thesaurus y de Medical Subject Headings que, aunque proporcionan una cobertura del contexto menor (70% y 43% respectivamente), cuentan con una estructura mucho más rica y son ontologías muy relevantes en la comunidad biomédica. La obtención de una cobertura del contexto máxima también tiene sus efectos en el ranking, y sitúa a la ontología en tercera posición, por encima de otras ontologías más populares y ricas que ella.

En cuanto a la salida combinada para este caso de estudio, ésta no se presenta, pues no aporta nueva información sobre lo que se pretendía demostrar con este experimento, respecto a la salida simple.

Tabla 7.10. Resultados del prototipo para el caso de estudio 4 con salida simple. Se resalta la información correspondiente a la ontología de ejemplo.

Posición	Ontología	CCscore	SRscore	Pscore	Score final
1	NCI Thesaurus	0,700	0,884	0,863	0,779
2	Medical Subject Headings	0,433	0,868	1,000	0,654
3	Ontología de ejemplo	1,000	0,024	0,000	0,556
4	Physician Data Query	0,200	0,768	0,703	0,443
5	MedlinePlus Health Topics	0,167	0,407	0,483	0,289
6	CRISP Thesaurus	0,300	0,322	0,180	0,283
7	Cancer Research and Management ACGT Master Ontology	0,200	0,173	0,323	0,216
8	Breast Tissue Cell Lines Ontology	0,133	0,239	0,280	0,189
9	Cell Line Ontology (MCCL)	0,133	0,239	0,280	0,189
10	-	-	-	-	-

7.2.6 Tiempos de ejecución

En este apartado se muestran los tiempos de ejecución del prototipo para cada caso de estudio (ver tabla 7.11), obtenidos con un equipo de características estándar⁷⁷. El tiempo para cada caso se ha calculado como la media de cinco ejecuciones sucesivas.

Durante el proceso de evaluación de cada ontología se realizan diversos accesos a ella para comprobar, por ejemplo, el número de términos iniciales que contiene, la cantidad de términos similares a los términos iniciales que figuran en ella, los parientes de cada concepto que cubre un término de entrada, etc. Esto se traduce en accesos a la BD que contiene las ontologías candidatas (repositorio de ontologías).

En la tabla 7.11 se diferencia entre el tiempo de ejecución obtenido cuando el caso de estudio se ejecuta por primera vez y entre el tiempo obtenido para las siguientes ejecuciones, hasta un reinicio del sistema. Cuando el caso de estudio se ejecuta por primera vez, ninguna de las consultas a la BD necesarias para evaluar la ontología se han ejecutado previamente y, por lo tanto, es necesario realizar múltiples accesos

⁷⁷ Intel Core 2 Quad 2.66GHz con 4 GB de memoria RAM.

“reales” a la BD. En las siguientes ejecuciones, los resultados de las consultas ejecutadas previamente se encuentran en la caché de BD. Los accesos a la caché de BD son notablemente más rápidos que los accesos a la propia BD. También se puede observar que los tiempos de ejecución más bajos se obtienen para los casos de estudio con menor número de términos iniciales. Esto se debe a que un menor número de términos iniciales implicará, generalmente, un menor número de sinónimos tras el proceso de expansión semántica y, por lo tanto, un menor número de consultas sobre el repositorio de ontologías durante el proceso de evaluación.

Tabla 7.11. Tiempos de ejecución medios (en segundos) tras 5 ejecuciones de los casos de estudio, para la salida simple y combinada (con combinaciones de 5 ontologías como máximo). Se muestra el número de términos iniciales y de ontologías candidatas para cada caso.

	Tiempos de ejecución (seg.)					
			Primera ejecución		Siguientes ejecuciones	
	Nº términos	Nº ontologías	Salida simple	Salida combinada	Salida simple	Salida combinada
CE 1	30	200	353,27	354,15	3,87	3,92
CE 2	32	200	181,31	182,18	7,03	8,08
CE 3	24	200	121,06	121,75	4,15	6,95
CE 4	80	200	621,92	631,15	7,28	22,34
CE 5	30	201	364,65	364,87	3,90	3,96

Los tiempos de ejecución obtenidos se pueden considerar razonables para un uso general del sistema de selección. Sin embargo, deberían ser mejorados si fuese necesario integrar el sistema en procesos que requieran reutilización de ontologías en tiempo real (e.g. algunas posibles aplicaciones de Web Semántica). Los tiempos obtenidos se podrían mejorar de diversas maneras: utilizando una máquina de mayor potencia y memoria; manteniendo una caché de BD de gran tamaño, que contenga un número elevado de posibles consultas; optimizando la configuración del sistema de gestión de BD (SGBD) utilizado; mejorando algunos algoritmos (e.g. búsqueda de conceptos similares a un concepto dado), etc. Sin embargo, estas tareas quedarían fuera del alcance de los objetivos de esta tesis, y se plantearán como un futuro trabajo.

7.3 Resultados de la evaluación

Para llevar a cabo la evaluación, se ha contactado por correo electrónico con varios expertos de prestigio internacional en el campo de las ontologías biomédicas, solicitándoles su colaboración para evaluar el trabajo. Se les ha pedido que cubriesen el *Formulario de evaluación por parte de expertos*, que se encuentra disponible en la dirección <http://tinyurl.com/5uobbnw>, y también en el anexo II de este documento.

Se ha recibido respuesta de 7 expertos. Cuatro de ellos han cubierto el formulario anteriormente mencionado, mientras que los 3 restantes han proporcionado su opinión y varias sugerencias a través de correo electrónico. En este apartado, se presentan los resultados obtenidos tras analizar las respuestas proporcionadas por los 4 expertos que han cubierto el cuestionario, completándolos con las recomendaciones proporcionadas por los otros 3 expertos.

Como se ha explicado en el apartado 7.1, el *Formulario de evaluación por parte de expertos* se compone de tres secciones: (1) Información personal. (2) Casos de estudio. (3) Preguntas generales. Este cuestionario recoge la opinión del experto acerca de las ontologías más adecuadas para cada caso de estudio, tanto para el caso de una única ontología (salida simple), como para la combinación de varias (salida combinada). También se le pide al experto que ejecute el sistema para cada caso, y que reflexione acerca de la salida proporcionada por el sistema respecto a la proporcionada por el experto de forma manual. Para los casos de estudio 1-4, el evaluador debe puntuar la utilidad del sistema de acuerdo a la siguiente escala (la puntuación asignada a cada opción se muestra entre paréntesis):

- Nada útil (0)
- Poco útil (1)
- Útil (2)
- Muy útil (3)

Esta escala también se ha utilizado en la sección 3 del cuestionario, para pedir al experto su opinión acerca de la utilidad del sistema, desde un punto de vista general. Para el caso de estudio 5, se ha preguntado a los expertos acerca de su opinión sobre la posición de la ontología de ejemplo en el ranking, de acuerdo a las siguientes opciones:

- La posición de la ontología de ejemplo en el ranking es adecuada.
- La posición de la ontología de ejemplo en el ranking no es adecuada, debería encontrarse más arriba.
- La posición de la ontología de ejemplo en el ranking no es adecuada, debería encontrarse más abajo.
- No sé si la posición es adecuada o no.

Los valores medios de las puntuaciones proporcionadas por los evaluadores para el sistema en los casos de estudio 1-4, así como para la utilidad general del sistema, se presentan en la tabla 7.12. Los resultados de la evaluación para el caso de estudio 5 se pueden ver en la tabla 7.13.

En el cuestionario se indica a los evaluadores que, en caso de no disponer de tiempo suficiente, podrían cubrir únicamente uno de los casos de estudio de la sección 2 del cuestionario en lugar de cubrir los 5. De esta manera, no todos los evaluadores han cubierto el mismo número de apartados del cuestionario. La columna “Nº respuestas” de la tabla 7.12 indica el número de evaluadores que han cubierto cada parte del cuestionario.

Tabla 7.12. Valores medios de las puntuaciones proporcionadas por los evaluadores para cada caso de estudio (casos 1-4) y para la opinión general sobre la utilidad del sistema. La columna “Nº respuestas” indica el nº de evaluadores que han cubierto cada caso. También se muestran los valores medios para el nº de respuestas y para cada tipo de salida.

Aspecto evaluado	Nº respuestas	Valores medios	
		Salida simple	Salida combinada
CE 1	4	2,50	2,25
CE 2	2	2,50	2,50
CE 3	2	3,00	2,50
CE 4	1	3,00	2,00
Valores medios	2,25	2,75	2,31
Utilidad general del sistema	4	2,75	2,25

Los resultados de la tabla 7.12 permiten ver que los evaluadores se han mostrado muy positivos acerca de la utilidad del sistema para seleccionar la mejor ontología u ontologías para un contexto determinado. Tanto para los casos de estudio 1-4 como desde un punto de vista general, los valores medios de las puntuaciones proporcionadas por los evaluadores se encuentran entre 2 (“Útil”) y 3 (“Muy útil”). En

cuanto al tipo de salida, consideran que la salida simple resulta de mayor utilidad que la salida combinada.

Por lo que respecta al caso de estudio 5 (ver tabla 7.13), todos los evaluadores que han cubierto la parte correspondiente a este caso de estudio consideran que el resultado obtenido es adecuado, es decir, están de acuerdo en que la ontología de ejemplo no debe figurar en la primera posición del ranking, pues no es una ontología consensuada por la comunidad biomédica, pero tampoco debe descender hasta las últimas posiciones, ya que cubre la totalidad de los términos de entrada.

Tabla 7.13. Resultados de la evaluación para el caso de estudio 5.

Respuesta para el CE 5	Nº ocurrencias
La posición es adecuada	3
Debería estar más arriba	0
Debería estar más abajo	0
No sabe	0
No contesta	1

En la tabla 7.14 se muestran otros aspectos de interés relativos a la evaluación realizada por los expertos, como el tiempo medio utilizado por los evaluadores para elaborar los rankings de forma manual, el tiempo medio que los evaluadores creen que necesitarían para elaborar un ranking ideal, etc.

Tabla 7.14. Otros aspectos de interés relativos al proceso de evaluación.

Aspecto	Resultado
Tiempo medio utilizado por los evaluadores para elaborar manualmente el ranking	Salida simple: 10 min. Salida combinada: 17 min.
Tiempo medio que creen que necesitarían los evaluadores para proporcionar manualmente un ranking ideal	Salida simple: 64 min. Salida combinada: 101 min.
Tiempo medio utilizado para cubrir el cuestionario (considerando sólo los evaluadores que han cubierto el cuestionario completo)	53 min.
Recursos de apoyo utilizados por los evaluadores para elaborar los rankings manualmente	BioPortal ⁷⁸ , OBO Foundry ⁷⁹ , Ontology Lookup Service ⁸⁰ , OntoBee ⁸¹ , Google

⁷⁸ <http://bioportal.bioontology.org/>

⁷⁹ <http://www.obofoundry.org/>

⁸⁰ <http://www.ebi.ac.uk/ontology-lookup/>

⁸¹ <http://www.ontobee.org/>

En la tabla 7.15 se resumen las principales ventajas e inconvenientes del sistema indicadas por los expertos, ordenadas según el número de evaluadores que las han mencionado. Como principal ventaja, destaca el hecho de que el sistema permite realizar la selección de ontologías de forma mucho más rápida y precisa que si se realiza de forma manual. Como se ha visto en la tabla 7.14, los tiempos medios estimados por los evaluadores para seleccionar las mejores ontologías en los casos de estudio superan los 60 minutos. Estos tiempos son muy superiores a los tiempos de respuesta del sistema, que se han mostrado en el apartado 7.2.6. El sistema permite obtener un resultado preciso de forma completamente automática en un tiempo considerablemente menor que si la tarea se realiza de forma manual. Los expertos también resaltan la gran utilidad del sistema en ámbitos con los que se encuentran poco familiarizados, o bien en campos que conocen en profundidad pero para los que existen muchas ontologías diferentes, entre las que resulta difícil elegir la más adecuada.

Tabla 7.15. Principales ventajas e inconvenientes del sistema, ordenadas según el número de evaluadores que las han mencionado. Nótese que, en este caso, se han tenido en cuenta 7 evaluadores (los 4 que han cubierto el cuestionario, más los 3 que han proporcionado su opinión por correo electrónico).

Ventajas	Nº evaluadores
Permite realizar el proceso de selección de ontologías de forma rápida y precisa	5
De gran utilidad en temas poco conocidos por el usuario, o bien en ámbitos conocidos, pero para los que existen muchas ontologías diferentes (e.g. anatomía)	5
Resulta de ayuda para confirmar la opinión del experto respecto a si una ontología es adecuada o no para un contexto	3
El sistema es fácil de usar, y los resultados obtenidos son fáciles de entender	2
El sistema indica los términos cubiertos y no cubiertos por las ontologías	1
Inconvenientes/Recomendaciones	
Proporcionar más instrucciones sobre el uso del sistema	3
Se debería indicar el grado de formalidad de cada fuente (ontología, vocabulario controlado, etc.)	3
Sería interesante disponer de la URI de cada ontología	1
Sería útil disponer de enlaces desde cada término a su definición y conceptos relacionados en la ontología	1
Utilizar directamente las ontologías del OBO Foundry como repositorio del sistema	1

Por otra parte, entre las recomendaciones indicadas por los evaluadores para mejorar el sistema se encuentra proporcionar más instrucciones en el sitio Web, que

faciliten el uso del sistema y la interpretación de los resultados, e indicar el grado de formalidad de cada ontología del repositorio, diferenciando entre aquellas ontologías menos formales o vocabularios controlados (e.g. NCI Thesaurus, Medical Subject Headings, etc.) de aquéllas de mayor formalidad (e.g. Gene Ontology).

Otras observaciones importantes, derivadas de la evaluación, son las siguientes:

- Para cada caso de estudio, se ha pedido a los expertos que indicasen un ranking de las ontologías más adecuadas para afrontarlo, teniendo en cuenta las 200 ontologías almacenadas en el repositorio. En cuanto a esta tarea, los expertos han proporcionado (en cuestión de minutos, como se puede ver en la tabla 7.14) varias ontologías que ellos considerarían candidatas para cada caso de estudio. Las ontologías elegidas por los expertos han sido, generalmente, ontologías muy conocidas en el dominio biomédico (e.g. NCI Thesaurus, Medical Subject Headings, Gene Ontology, etc.), o bien ontologías no tan populares, pero que cada experto conoce bien debido a su experiencia personal (e.g. National Drug File, Cell Line Ontology, etc.). Los expertos han reconocido que desconocían algunas de las ontologías contempladas por el sistema y coinciden en que las ontologías que han proporcionado manualmente y en pocos minutos, son únicamente un filtrado inicial e impreciso para afrontar cada caso de estudio. Indican que una selección en más profundidad, que permitiese resolver adecuadamente el caso de estudio, les llevaría mucho más tiempo (más de 1 hora, de media, como se puede ver en la tabla 7.14). Señalan que con más tiempo, podrían familiarizarse con las ontologías que no conocen, y también revisar qué ontología contiene más términos de entrada. Los expertos también indican que, tras observar los resultados proporcionados por el sistema para cada caso de estudio, reordenarían los rankings elaborados por ellos en base a la cantidad de términos cubiertos por cada ontología, e incluirían nuevas ontologías no contempladas por ellos inicialmente.
- Los expertos han considerado que la salida simple resulta de mayor utilidad de la salida combinada. Sin embargo, consideran que la salida combinada resulta útil, y son conscientes de la considerable cantidad de tiempo que se necesita para obtener una salida combinada de forma manual (101 minutos de media para los casos de estudio).

- Los resultados obtenidos para el caso de estudio 5 han mostrado la importancia de combinar tres criterios independientes, pero complementarios, para alcanzar resultados sólidos en el proceso de selección. Todos los evaluadores que han cubierto el apartado del cuestionario relativo a este caso de estudio han considerado que el funcionamiento del prototipo para afrontar dicha tarea ha sido adecuado.
- La tabla 7.16 permite ver las ontologías que han aparecido más de una vez en los 10 primeros resultados proporcionados por el prototipo, para los casos de estudio 1 al 4. Se ha obviado el caso de estudio 5 por ser redundante con el caso de estudio 1. Se puede ver que un 70% (7 de 10) de las ontologías que figuran en esta tabla son ontologías de gran tamaño, con más de 15.000 conceptos. El gran número de conceptos en estas ontologías facilita que proporcionen mejores resultados de cobertura del contexto que ontologías de menor tamaño.

Tabla 7.16. Ontologías seleccionadas por el sistema más de una vez en el “top 10”, para los casos de estudio 1-4. Se resaltan las ontologías que han sido seleccionadas 3 veces o más.

Ontología	Nº conceptos	Nº apariciones
NCI Thesaurus	67.803	4
Medical Subject Headings	295.830	4
CRISP Thesaurus	16.682	4
Logical Observation Identifier Names and Codes	114.351	3
Alcohol and Other Drug Thesaurus	15.888	2
Experimental Factor Ontology	3.157	2
Cardiac Electrophysiology Ontology	81.697	2
Cancer Research and Management ACGT Master Ontology	1.770	2
Galen Ontology	23.141	2
HL7 Vocabulary Version 2.5	4.911	2

- La ontología sobre cáncer NCI Thesaurus ocupa las primeras posiciones del ranking para todos los casos de estudio, aún a pesar de no ser la ontología que contiene más conceptos, ni la más popular, y teniendo en cuenta que 2 de los 5 casos de estudio no se centran en el dominio del cáncer. También destaca el hecho de que la ontología CRISP Thesaurus ha sido seleccionada 4 veces en el “top 10”, tantas como el NCI Thesaurus o Medical Subject Headings, a pesar de contener un número de conceptos mucho menor (16.000, respecto a

295.000 de Medical Subject Headings). La ontología CRISP (Computer Retrieval of Information on Scientific Projects) ha sido creada por el Instituto de Salud Americano (NIH), y contiene términos relativos a los programas y proyectos de investigación biomédica financiados por el gobierno de Estados Unidos.

8 Conclusiones y futuras líneas de investigación

En este capítulo, se presentan las principales conclusiones a las que se ha llegado tras la realización de este trabajo. Se concluye apuntando algunas posibles líneas de investigación.

8.1 Conclusiones

Durante los últimos años, se han desarrollado múltiples ontologías del dominio biomédico, que suponen un elemento crucial para la representación del conocimiento y la anotación de datos en este dominio. Reutilizar el conocimiento de estas ontologías para abordar nuevos problemas en lugar de crear ontologías nuevas, con conocimiento redundante, es una práctica esencial para lograr una adecuada interoperabilidad. Sin embargo, la gran cantidad y complejidad de las ontologías biomédicas existentes, en continua evolución, hacen cada vez más difícil la tarea de seleccionar la ontología u ontologías más adecuadas para un determinado contexto o problema y, por lo tanto, esto supone una importante barrera para su reutilización.

En este trabajo, se ha presentado una aproximación para la selección automática de ontologías biomédicas, que se basa en evaluar cada ontología candidata de acuerdo a tres criterios: (1) El grado de cobertura del contexto proporcionado por la ontología. (2) La riqueza semántica de la ontología en el contexto. (3) La popularidad de la ontología, calculada de acuerdo al conocimiento colectivo almacenado en varios recursos Web 2.0. Para validar la aproximación, se ha implementado un prototipo de sistema de selección automática de ontologías biomédicas, que se ha bautizado como

BiOSS (Biomedical Ontology Selection System), y que se encuentra disponible públicamente en la dirección <http://bioss.ontologyselection.com>. Con la ayuda de varios expertos, se ha comprobado el correcto funcionamiento del prototipo y la validez de la aproximación para resolver problemas reales. Los experimentos realizados demuestran empíricamente que la hipótesis planteada en el apartado 1.4 de esta tesis es cierta, haciendo posible concluir que: combinando tres criterios de evaluación independientes, pero complementarios, como son el grado de cobertura del contexto proporcionado por la ontología, la riqueza semántica de su estructura y elementos, y su popularidad o relevancia en la comunidad biomédica, es posible idear una aproximación para la adecuada selección automática de ontologías en el ámbito biomédico.

Otras de las conclusiones más relevantes derivadas de la realización de este trabajo, son las siguientes:

- Desarrollar ontologías es una tarea difícil y costosa, que depende de una gran variedad de requerimientos y expectativas de los usuarios. Debido a esto, es importante disponer de aproximaciones y herramientas automáticas que permitan seleccionar las ontologías más adecuadas para describir un dominio particular, y así poder reutilizarlas.
- El núcleo de la selección de ontologías es la evaluación de ontologías, y es en la evaluación de ontologías donde debe centrarse el esfuerzo si se desea obtener buenos resultados de selección.
- La popularidad de las ontologías es un criterio a tener en cuenta para obtener unos resultados adecuados en el proceso de selección. Además, se ha visto que el conocimiento colectivo almacenado en recursos Web constituye una fuente adecuada para este tipo de evaluación, sin necesidad de obligar a los usuarios a realizar un esfuerzo de valoración manual de cada ontología.
- La importancia de las aproximaciones y herramientas de evaluación y selección de ontologías continuará aumentando a medida que continúen desarrollándose ontologías de diferentes dominios y repositorios de ontologías que las almacenen, pero alcanzarán su máximo potencial en el momento en que existan

agentes de Web Semántica con una necesidad real de disponer de mecanismos automáticos de reutilización de conocimiento.

- La mejor forma de saber qué contiene una ontología y cómo de adecuada resulta para describir un dominio es examinar su contenido o estructura. A medida que el número de ontologías vaya aumentando, irán apareciendo nuevos repositorios de ontologías que incorporarán metadatos sobre ellas, describiendo tanto aspectos objetivos como subjetivos de las mismas. Esto puede facilitar considerablemente las tareas de evaluación de ontologías. Sin embargo, estos metadatos deberán ser tratados con cautela, pues pueden ser erróneos. Durante la evolución de la Web, algunos motores de búsqueda (e.g. Google) decidieron no utilizar los metadatos de las páginas Web (e.g. palabras clave), porque los autores de las páginas los utilizaban de forma estratégica para mejorar el posicionamiento en los buscadores. Es posible que esto también ocurra con las ontologías.

8.2 Futuras líneas de investigación

A continuación, se plantean algunas líneas de investigación, derivadas o directamente relacionadas con el trabajo desarrollado en esta tesis, en las que puede resultar de interés profundizar durante los próximos años:

- La anotación semántica es un proceso difícil, para el que no existen actualmente herramientas adecuadas. Una posible futura línea de trabajo consistiría en construir un sistema de anotación semántica de recursos biomédicos basado en el sistema de selección de ontologías propuesto. Se podría adecuar la entrada del sistema de selección para la anotación semántica de diferentes tipos de recursos (e.g. textos, artículos científicos, imágenes, etc.).
- Otra línea que podría resultar de gran interés es la aplicación de técnicas de razonamiento basado en ontologías a las áreas de la evaluación y selección de ontologías. La adecuada aplicación de estas técnicas podría permitir extraer información sobre las ontologías relevante para su evaluación, muy difícil o imposible de obtener usando las actuales técnicas.

- Un campo en que ya se está trabajando (principalmente en el ámbito biomédico), y que resultará de gran interés durante los próximos años, será el relacionado con la construcción y mantenimiento de repositorios de ontologías a gran escala. En este ámbito existen diversos aspectos que pueden resultar de interés para comenzar una investigación. Algunos ejemplos son el desarrollo de agentes semánticos inteligentes que exploren la Web en busca de ontologías de calidad para incorporarlas al repositorio, el estudio de la detección y actualización de las correspondencias inter-ontología (*mappings*), la detección de errores o inconsistencias en las ontologías contenidas en el repositorio, etc.
- Otra posible línea de trabajo consistiría en estudiar la forma de mejorar el tiempo de ejecución del prototipo implementado. Un aspecto que irá adquiriendo importancia a medida que la Web Semántica vaya imponiéndose, será investigar la forma de incrementar el tiempo de respuesta de las técnicas de evaluación y selección de ontologías. En un contexto de Web Semántica, será necesario disponer de sistemas que permitan abordar tareas de reutilización automática de conocimiento de forma instantánea.
- Utilizar el prototipo que se ha implementado para estudiar las actuales ontologías biomédicas, analizando la correlación entre diferentes variables: cobertura del contexto, riqueza semántica, popularidad, número de conceptos, fecha de publicación, etc. Con esto se pretende dar respuesta a preguntas como: ¿las ontologías que mejor cubren el contexto son las más populares?, ¿la cantidad de conceptos de la ontología influye en su riqueza semántica?, ¿y en su popularidad?, ¿existen ontologías muy “jóvenes” que sean muy populares?, etc.

9 Conclusions and future research lines

This chapter presents the main conclusions reached after the completion of this work. In addition, some possible research lines are proposed.

9.1 Conclusions

In recent years, many ontologies for the biomedical domain have been developed, which constitute an essential element for knowledge representation and data annotation in such domain. Reusing knowledge from these ontologies in order to solve new problems, instead of building new ontologies, with redundant knowledge, is a common practice to ensure a proper interoperability. However, the large number and high complexity of current biomedical ontologies, which continues evolving, makes it more and more difficult the task of selecting the ontology or ontologies more adequate for a given context or problem and, due to this, these issues constitute an important barrier to reuse them.

In this work, an approach for the automatic selection of biomedical ontologies has been proposed, which is based on evaluating each candidate ontology according to three criteria: (1) The level of context coverage provided by the ontology. (2) The semantic richness of the ontology. (3) The popularity of the ontology, which is calculated on the basis of collective knowledge extracted from several Web 2.0 resources. In order to validate the approach, a system for the automatic selection of biomedical ontologies has been prototyped, which is known as BiOSS (Biomedical Ontology Selection System), and which is publicly available at <http://bioss.ontologyselection.com>. With the support of several experts, the

correctness of the prototype and the validity of the approach to solve real problems has been demonstrated. Experiments have empirically shown that the hypothesis proposed in section 1.4 of this document is true, making it possible to conclude that: combining three independent, but complementary evaluation criteria, it is possible to devise an approach for the adequate selection of ontologies in the biomedical domain.

Others of the most important conclusions of this work are the following:

- Ontology development is a difficult and time-consuming task, which depends on a large variety of users' requirements and expectations. Due to this, it is important to have automatic approaches and tools to select the most adequate ontologies to describe a particular domain, in order to reuse them.
- The core of the ontology selection process is ontology evaluation, and this is the task that should focus the effort, in order to obtain good selection results.
- Popularity of ontologies is an important criterion to obtain adequate results in the selection process. In addition, collective knowledge stored in Web resources is an adequate source for this kind of evaluation, without forcing users to perform a manual evaluation effort of each ontology.
- The importance of approaches and tools for ontology evaluation and selection will continue to increase as ontologies for different domains continue being developed, as well as ontology repositories to store them. However it will reach its full potential when there exist Semantic Web agents with a real need of mechanisms for automatic knowledge reusing.
- The best way of knowing what is contained in an ontology and how appropriate is to describe a domain, is examining its content or structure. As the number of ontologies increases, new ontology repositories will arise. These repositories will store diverse metadata for ontologies, describing both objective and subjective ontological aspects. This will probably facilitate in a considerable manner the ontology evaluation tasks. However, these metadata should be treated with caution, as they may be wrong. During the evolution of the Web, some search engines (e.g. Google) decided not to use metadata from webpages (e.g. keywords), because the authors of the pages started to use them

strategically to improve their results in search rankings. This may also happen to ontologies.

9.2 Future research lines

In the following, some research lines are proposed, arising from or directly related to the work developed in this thesis, which may be of interest to study in the following years:

- Semantic annotation is a difficult process, for which there are currently no appropriate tools. A possible future line of work would consist on building a system for the semantic annotation of biomedical resources, based on the ontology selection system that has been proposed. The input of the system could be adjusted to the semantic annotation of different kind of resources (e.g. texts, scientific articles, images, etc.).
- Another line that could be of great interest is the application of ontology based reasoning techniques to the areas of ontology evaluation and selection. The appropriate application of these techniques could allow to extract information about ontologies that could be used during the evaluation, and that would be very difficult or impossible to obtain by means of current techniques.
- A field in which researchers are already working (mainly in the biomedical domain), and that will be of great interest over the next years, will be related to the construction and maintenance of large-scale ontology repositories. In this area, there are several aspects that may be of interest to begin a research. Some examples are the development of intelligent semantic agents that explore the Web looking for high-quality ontologies, in order to incorporate them to the repository; the study of identification and update of inter-ontology mappings; detection of errors or inconsistencies for ontologies in the repository, etc.
- Another possible line of work is to study how to improve execution time of the implemented prototype. One issue which will become more important as the Semantic Web will be imposed, will be to research the way of increasing response time of evaluation and selection techniques. In a Semantic Web

context, it will be necessary to have systems to solve tasks of automatic knowledge reusing instantly.

- Using the prototype that has been implemented to study current biomedical ontologies, by analyzing the correlation between different variables: context coverage, semantic richness, popularity, number of concepts, publication date, etc. This is intended to answer questions such as: are the ontologies that best cover the context the most popular?, do the amount of concepts in the ontology influence its semantic richness?, and its popularity?, are there any ontologies that are “young” and popular at the same time?, etc.

Acrónimos

A continuación, se listan los acrónimos utilizados en este documento.

ALA	American Library Association
API	Application Programming Interface
ASA	Abstracción Estructural Anatómica
ATA	Abstracción de Transformación Anatómica
ARPA	Advanced Research Projects Agency
BD	Base de Datos
BEM	Betweenness Measure
BiOSS	Biomedical Ontology Selection System
CC:DA	Committee on Cataloging: Description and Access
CCscore	Context Coverage Score
CE	Caso de Estudio
CERN	European Organization for Nuclear Research
CIE	Clasificación Internacional de Enfermedades
CMM	Class Match Measure
CUI	Concept Unique Identifier
DAML	DARPA Agent Markup Language
DAO	Data Access Object
DEM	DEnsity Measure
DL	Description Logic
DVD	Digital Versatile Disc
EBI	European Bioinformatics Institute

EHR	Electronic Health Records
FMA	Foundational Model of Anatomy
FTP	File Transfer Protocol
GO	Gene Ontology
GUI	Graphic User Interface
HTML	HyperText Markup Language
IA	Inteligencia Artificial
ICD	International Classification of Diseases
IHTSDO	International Health Terminology Standards Development Organization
IMEDIR	Centro de Informática Médica y Diagnóstico Radiológico
ISBN	International Standard Book Number
KBS	Knowledge Based System
KIF	Knowledge Interchange Format
KR	Knowledge Representation
KSL	Knowledge Systems Laboratory
KST	Knowledge Sharing Technology
LCA	Lowest Common Ancestor
MCST	Most Central Semantic Type
MeSH	Medical Subject Headings
MGD	Mouse Genome Database
NIH	National Institutes of Health
NLM	National Library of Medicine
NLP	Natural Language Processing
NCBO	National Center for Biomedical Ontology

NCI	National Cancer Institute
NCICB	National Cancer Institute Center for Bioinformatics
NLM	National Library of Medicine
NLP	Natural Language Processing
NS	NameSpace
OBO	Open Biomedical Ontologies
OCML	Operational Conceptual Modeling Language
OIL	Ontology Inference Layer
OKBC	Open Knowledge Base Connectivity
OMIM	Online Mendelian Inheritance in Man
OMS	Organización Mundial de la Salud
ORS	Open Rating System
OWL	Ontology Web Language
PC	Personal Computer
PDF	Portable Document Format
Pscore	Popularity Score
RAE	Real Academia Española
RDF	Resource Description Framework
RDF(S)	Resource Description Framework Schema
RRF	Rich Release Format
RSS	Really Simple Syndication
SBC	Sistema Basado en Conocimiento
SGBD	Sistema de Gestión de Base de Datos
SGD	Saccharomyces Genome Database
SGML	Standard Generalized Markup Language

SHOE	Simple HTML Ontology Extensions
SNOMED-CT	Systematized Nomenclature of Medicine - Clinical Terms
SNOMED-RT	Systematized Nomenclature of Medicine - Reference Terminology
SOAP	Simple Object Access Protocol
SPARQL	SPARQL Protocol and RDF Query Language
SQL	Structured Query Language
SRscore	Semantic Richness Score
SSM	Semantic Similarity Measure
UCI	Unidad de Cuidados Intensivos
UDC	Universidade da Coruña
UKACR	United Kingdom Association of Cancer Registries
UMLS	Unified Medical Language System
UNSPSC	United Nations Standard Products and Services Code
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
VO	Value Object
W3C	World Wide Web Consortium
WSD	Word Sense Disambiguation
WSDL	Web Services Description Language
WWW	World Wide Web
XML	eXtensible Markup Language
XOL	XML-Based Ontology Exchange Language

Anexo I. Ontologías utilizadas

A continuación, se muestran los nombres y las versiones de las 200 ontologías que componen el repositorio creado para este trabajo.

Ontología	Versión
ABA Adult Mouse Brain	1.0
Adverse Event Ontology	1.0.23
African Traditional Medicine Ontology	1.101
AI/RHEUM	1993
Alcohol and Other Drug Thesaurus	2000
Amino Acid Ontology	1.2 (inferred)
Amphibian Gross Anatomy Ontology	1.8
Amphibian Taxonomy	26/02/2009
Animal Diversity Web Ontology	10/11/2004
Apollo-akesios ontology	0.1
Ascomycete phenotype ontology	1.16
Authorized Osteopathic Thesaurus	2003
Basic Formal Ontology	1.1.1
Basic Vertebrate Anatomy Ontology	1.1
Bilateria Anatomy Ontology	16/07/2008
Biomedical Resource Ontology	3.2.1
BioPAX Ontology	Level3 v1.0
BioPortal Metadata	0.9.3
BioTop Ontology	dev
BIRNLex	1.3.1
Bleeding History Phenotype Ontology	0.3.1
Body System Ontology	1.0
Breast Tissue Cell Lines Ontology	2.0
BRENDA tissue / enzyme source	02/11/2010
Brucellosis Ontology	1.0.67
C. elegans development ontology	1.3
C. elegans phenotype ontology	1.159
Cancer Research and Management ACGT Master Ontology	1.1
Cardiac Electrophysiology Ontology	1.0
Cell Behavior Ontology	0.0
Cell Line Ontology (CLO)	1.0
Cell Line Ontology (MCCL)	1.0
Cell Type Ontology	1.48
Cereal Plant Development Ontology	1.6
Cereal Plant Gross Anatomy Ontology	unknown
Cereal Plant Trait Ontology	unknown
Chemical Entities of Biological Interest Ontology	unknown
Chemical Information Ontology	unknown
Clinical Classifications Software	2005
Common Anatomy Reference Ontology	1.5
Common Terminology Criteria for Adverse Events	4.02
Comparative Data Analysis Ontology	1.26
Computer-based Patient Record Ontology	0.85
Computer-Stored Ambulatory Records	89-95
COSTART	1995
CRISP Thesaurus	2006
Dendritic Cell Ontology	unknown
DermLex: The Dermatology Lexicon	1.0

Dictyostelium Discoideum Anatomy Ontology	1.10
Drosophila Development Ontology	1.17
Drosophila Gross Anatomy Ontology	1.39
DXplain	1994
EDAM Ontology	beta08
Electrocardiography Ontology	0.1.7
Environment Ontology	02/03/2010
Epoch Clinical Trial Ontologies	0.9
Event (INOH pathway ontology)	1.71
Evidence Codes Ontology	18/03/2010
eVOC (Expressed Sequence Annotation for Humans)	unknown
Experimental Factor Ontology	133
Family Health History Ontology	1.0
Fly Taxonomy	unknown
FlyBase Controlled Vocabulary	1.19
Foundational Model of Anatomy	2_0
Fungal Gross Anatomy Ontology	1.4
Galen Ontology	1.1
Gene Ontology	2008_04_01
Gene Regulation Ontology	0.5
General Formal Ontology	1.0
General Formal Ontology: Biology	1.1
GeoSpecies Ontology	beta
Healthcare Common Procedure Coding System	2009
HL7 Vocabulary Version 2.5	2003_08_30
HL7 Vocabulary Version 2.5, 7-bit equivalents	2003_08
HL7 Vocabulary Version 3.0	2006_05
HOM Hospital Discharge Codes	1.1
HUGO Gene Nomenclature	2008_03
Human developmental anatomy, abstract version	1.3
Human developmental anatomy, timed version	1.3
Human Phenotype Ontology	13/11/2010
Hymenoptera Anatomy Ontology	06/11/2009
ICD-10-PCS	2008
ICD-9-CM	2009
IMGT Ontology	1.0.0
Infectious Disease Ontology	18/11/2010
Information Artifact Ontology	15/07/2009
Interaction Network Ontology	1.0.17
International Classification for Nursing Practice	2
International Classification of Functioning, Disability and Health	1.0
International Classification of Primary Care	1993
Kinetic Simulation Algorithm Ontology	24/01/2009
Library of Congress Subject Headings	1990
linkingkin2pep Ontology	0.1
Lipid Ontology	18/11/2010
Loggerhead Nesting Ontology	unknown
Logical Observation Identifier Names and Codes	226
MaHCO - An MHC Ontology	1.0.1
Maize Gross Anatomy Ontology	1.1
Malaria Ontology	1.2.2
Mammalian Phenotype Ontology	11/11/2010
Mass Spectrometry Ontology	unknown
McMaster University Epidemiology Terms	1992
MDSS Mo Ontology	1.0
Medaka Fish Anatomy and Development Ontology	1.1
Medical Subject Headings	2009_2009_02_13
MedlinePlus Health Topics	20080614
MeGO Ontology	1.8
Metathesaurus additional entry terms for ICD-9-CM	2009
Metathesaurus CPT Hierarchical Terms	2009
Metathesaurus FDA National Drug Code Directory	2008_12_03

Metathesaurus FDA Structured Product Labels	2009_01_26
Metathesaurus HCPCS Hierarchical Terms	2009
Metathesaurus Source Terminology Names	
Metathesaurus Version of Minimal Standard Terminology Digestive Endoscopy	2001
MGED Ontology	1.3.1.1
Minimal Anatomical Terminology	1.1
Mosquito Gross Anatomy Ontology	1.10
Mosquito Insecticide Resistance Ontology	1.991
Mouse Adult Gross Anatomy Ontology	1.205
Mouse Pathology Ontology	1.4
Multiple Alignment Ontology	1.1
NanoParticle Ontology	2010-10-31 (inferred)
NCBI Taxonomy	2008_05_07
NCI Thesaurus	2008_05D
NeoMark Oral Cancer-Centred Ontology	3.1
Neural ElectroMagnetic Ontologies	1.30
Neural Motor Recovery Ontology	0.1
Neural-Immune Gene Ontology	1
NIF Cell Ontology	19/11/2010
NIF Dysfunction Ontology	19/11/2010
NMR-instrument specific component of metabolomics investigations	19/11/2010
OBOE Ontology	1.0
OBOE SBC Ontology	0.1
Online Mendelian Inheritance in Man	2007_12_19
Ontology for Biomedical Investigations	1.0
Ontology for disease genetic investigation	0.7.1
Ontology for Drug Discovery Investigations	0.9
Ontology for General Medical Science	20/07/2010
Ontology for Genetic Interval	0.15
Ontology for MicroRNA Target Prediction	0.2
Ontology of Clinical Research	Revision 120
Ontology of Geographical Region	1.1
Ontology of Glucose Metabolism Disorder	1.2
Ontology of Homology and Related Concepts in Biology	1.4
Ontology of Language Disorder in Autism	15/12/2008
Ontology of Physics for Biology	0.9.1.0
Parasite Experiment Ontology	0.20
Parasite Life Cycle Ontology	0.10
Pathogen Transmission Ontology	1.11
Pathway Ontology	1.060710
Perioperative Nursing Data Set, 2nd edition	2002
PHARE Ontology	21/10/2010
Phenotypic Quality Ontology	1.273
PhysicalFields Ontology	0.1
Physician Data Query	2007_02
Physico-chemical Process Ontology	1.13
Pilot Ontology	0.1
PKO_Re Ontology	1.1
Plant Environmental Conditions Ontology	1.6
Plant Growth and Developmental Stage Ontology	1.34
Plant Structure Ontology	1.34
Platynereis Stage Ontology	21/04/2010
PMA 2010 Ontology	0.9.1
Protein Ontology	2.0
Proteomics data and process provenance Ontology	1.1
Proteomics Pipeline Infrastructure for CPTAC	1.0
Pseudogene Ontology	0.1
QMR clinically related terms from Randolph A. Miller	1999
Quick Medical Reference (QMR)	1996
Rat Strain Ontology	2.1
Reproductive Traits Ontology	1.0
RNA Ontology	r113

Role Ontology	1.0.6
RxNorm Vocabulary	08AB_090202F
Sample Processing and Separation Techniques Ontology	1.070708
SemanticScience Integrated Ontology	alpha
Situation-Based Access Control Ontology	1.2
Skin Physiology Ontology	2.0
Smoking Behavior Risk Ontology	1.0
SNP-Ontology	1.6
Software Ontology	102
Spider Ontology	1.19
Standard Product Nomenclature	2003
Subcellular Anatomy Ontology	1.2.5
Symptom Ontology	18/11/2010
SysMO-JERM Ontology	0.6 Alpha
Systems Biology Ontology	1.0
Teleost Anatomy Ontology	1.198
Terminology for the Description of Dynamics	rel-2009-10-16 (inferred)
TEST Ontology	1.0
Tissue Microarray Ontology	1.0.0
TOK Ontology	0.2.1
Translational Medicine Ontology	r18
TSI Hospital Discharge Codes	1.1
University of Washington Digital Anatomist	1.7.3
USP Model Guidelines	2004
Vaccine Ontology	456
Veterans Health Administration National Drug File	2008_01_06
Wheat Trait Ontology	08/11/2010
Yeast Phenotypes Ontology	1.20

Anexo II. Formularios

Formulario de ajuste de pesos

- QUESTIONNAIRE FOR ADJUSTMENT OF WEIGHTS -

Name and Surname: _____ Date: ____ / ____ / ____

Supposing that it is necessary to select the right ontology from a set of biomedical ontologies, in order to semantically describe several input terms...

1. Distribute 100 points among the following criteria for evaluating biomedical ontologies, according to their relevance, or check this box [] if you consider that all the criteria have the same importance.

Criterion	Explanation	Points*
Context coverage	Amount of input terms (key words) contained in the ontology	
Semantic richness	Richness of the ontological structure	
Popularity	Relevance of the ontology in the Web	

2. Distribute 100 points among the following criteria for evaluating the semantic richness of biomedical ontologies, according to their relevance, or check this box [] if you consider that all the criteria have the same importance

Criterion	Explanation	Points*
Relatives	Number of direct relatives (i.e. parents, children and siblings) for each concept	
Additional information	Other knowledge in the ontology that can be of interest for a user (e.g. definitions, synonyms, comments, etc.)	
Similar knowledge	Other concepts contained in the ontology, similar to those concepts that describe the initial terms provided by the user	

3. Distribute 100 points among the following Web resources for evaluating the popularity of biomedical ontologies, according to their relevance, or check this box [] if you consider that all the resources criteria have the same importance.

Resource	Points*
Wikipedia	
Twitter	
BioPortal	
PubMed	

* Remember that the sum of the points in each table must be 100.

Formulario de evaluación por parte de expertos

Evaluation of BiOSS system

BiOSS is available at <http://bioss.ontologyselection.com>

The questionnaire has the following structure:

Section 1 - Personal information (4 questions)

Section 2 - Test cases

- Test case 1: Breast cancer (9 questions)
- Test case 2: Critical care (9 questions)
- Test case 3: UK cancer registries (9 questions)
- Test case 4: Anatomy (9 questions)
- Test case 5: Test case 1 adding an "Example ontology" (2 questions)

Section 3 - General questions (4 questions)

NOTE: If you don't have time to fill out all the test cases in Section 2, please, fill out at least one of them.

Thank you very much for your help.

[Continue »](#)

Powered by [Google Docs](#)

[Report Abuse](#) - [Terms of Service](#) - [Additional Terms](#)

Evaluation of BiOSS system

* Required

Section 1 - Personal information

Name and surname

Institution
University, research center, enterprise, etc.

E-mail address

Country *

[« Back](#) [Continue »](#)

Powered by [Google Docs](#)

[Report Abuse](#) - [Terms of Service](#) - [Additional Terms](#)

Evaluation of BiOSS system

Section 2 - Test cases

- Test case 1: Breast cancer (9 questions)
- Test case 2: Critical care (9 questions)
- Test case 3: UK cancer registries (9 questions)
- Test case 4: Anatomy (9 questions)
- Test case 5: Test case 1 adding an "Example ontology" (2 questions)

NOTE: If you don't have time to fill out all the test cases in this Section, please, fill out at least one of them.

[« Back](#) [Continue »](#)

Powered by [Google Docs](#)

[Report Abuse](#) - [Terms of Service](#) - [Additional Terms](#)

Evaluation of BiOSS system

Section 2 - Test case 1 (Breast cancer)

(if you do not want to fill out this test case, click on the "Continue" button at the bottom of this page)

DESCRIPTION: Researchers at Stanford University are building a system that abstracts clinical information from two electronic medical record databases related to the care and management of breast cancer. To build this application, this set of researchers need to reuse ontologies already developed by other organizations.

TERMS REPRESENTING THE DATA (30 terms): ductal carcinoma in situ, adjuvant chemotherapy, axillary lymph node staging, mastectomy, tamoxifen, serotonin reuptake inhibitors, invasive breast cancer, hormone receptor positive breast cancer, ovarian ablation, premenopausal women, surgical management, biopsy of breast tumor, fine needle aspiration, sentinel lymph node, breast preservation, adjuvant radiation therapy, prechemotherapy, inflammatory breast cancer, ovarian failure, bone scan, lumpectomy, brain metastases, pericardial effusion, aromatase inhibitor, postmenopausal, palliative care, guidelines, stage iv breast cancer disease, trastuzumab, breast mri examination.

1) Suppose that you have to choose ONE ontology (you can find a reference list of biomedical ontologies at <http://tinyurl.com/6dguy8u>) to solve this task. Please, provide a ranking (fill all positions as you can) of the ontologies you consider are the most adequate. You are allowed to use any website o tool you consider helpful.

Example: 1st: Gene Ontology; 2nd: NCI Thesaurus; 3rd: Medical Subject Headings, etc.

2) How much time did you take to answer the previous question? If you did not answer it, please, indicate why (e.g. I would need at least 1 hour, it is a too complex task, etc.)

3) Now, suppose that you are allowed to choose ONE OR SEVERAL biomedical ontologies to solve the task. Please, provide a ranking (fill all positions as you can) of the ontologies you consider are the most adequate. You are allowed to use any website o tool you consider helpful.

Example: 1st: Gene Ontology + NCI Thesaurus; 2nd: Gene Ontology + Medical Subject Headings + Galen Ontology; 3rd: Foundational Model of Anatomy; etc.

4) How much time did you take to answer the previous question? If you did not answer it, please, indicate why (e.g. I would need at least 1 hour, it is a too complex task, etc.)

5) If you had more time, do you think that the rankings you have provided would be better? How much time do you think you would need?

6) Have you used any tool or website to answer the previous questions? Which one/s?

7) Please, go to the BiOSS website (<http://bioss.ontologyselection.com>), choose "Test case 1 (Breast cancer)" as input and execute it (button "Go!") both in "Single Output" and "Combined Output" modes. Examine the top 10 results provided by the system. How helpful do you think that the results are to the proposed task?

	Not helpful	Few helpful	Helpful	Very helpful
Single output	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Combined output	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

8) After examining the results provided by the system, would you modify the rankings provided by you? How would you modify them?

9) Any other comments about the results provided by the system?

[« Back](#) [Continue »](#)

Powered by [Google Docs](#)

[Report Abuse](#) - [Terms of Service](#) - [Additional Terms](#)

Evaluation of BiOSS system

Section 2 - Test case 2 (Critical care)

(if you do not want to fill out this test case, click on the "Continue" button at the bottom of this page)

DESCRIPTION: Researchers at the University of A Coruña (Spain) need to build an ontology related to the monitorization at Intensive Care Units (ICUs) of patients who underwent heart surgery. In order to build such ontology, this group of researchers want to reuse knowledge from existing ontologies.

TERMS REPRESENTING THE DATA (32 terms): vital signs, mean arterial pressure, cardiac output, pressure venous central, expert systems, decision, serum, quadruple, add, stop, double, ascend, descend, continuous, medical device, pump, monitoring device, pharmacologic substance, vasodilator, nitroprusside, nitroglycerin, adrenaline, inotropic agents, noradrenaline, dobutamine, accepted, lower limit value, upper limit value, value, name, priority, variation.

1) Suppose that you have to choose ONE ontology (you can find a reference list of biomedical ontologies at <http://tinyurl.com/6dguy8u>) to solve this task. Please, provide a ranking (fill all positions as you can) of the ontologies you consider are the most adequate. You are allowed to use any website o tool you consider helpful.

Example: 1st: Gene Ontology; 2nd: NCI Thesaurus; 3rd: Medical Subject Headings, etc.

2) How much time did you take to answer the previous question? If you did not answer it, please, indicate why (e.g. I would need at least 1 hour, it is a too complex task, etc.)

3) Now, suppose that you are allowed to choose ONE OR SEVERAL biomedical ontologies to solve the task. Please, provide a ranking (fill all positions as you can) of the ontologies you consider are the most adequate. You are allowed to use any website o tool you consider helpful.

Example: 1st: Gene Ontology + NCI Thesaurus; 2nd: Gene Ontology + Medical Subject Headings + Galen Ontology; 3rd: Foundational Model of Anatomy; etc.

4) How much time did you take to answer the previous question? If you did not answer it, please, indicate why (e.g. I would need at least 1 hour, it is a too complex task, etc.)

5) If you had more time, do you think that the rankings you have provided would be better? How much time do you think you would need?

6) Have you used any tool or website to answer the previous questions? Which one/s?

7) Please, go to the BiOSS website (<http://bioss.ontologyselection.com>), choose "Test case 2 (Critical care)" as input and execute it (button "Go!") both in "Single Output" and "Combined Output" modes. Examine the top 10 results provided by the system. How helpful do you think that the results are to the proposed task?

Not helpful Few helpful Helpful Very helpful

Single output	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Combined output	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

8) After examining the results provided by the system, would you modify the rankings provided by you? How would you modify them?

9) Any other comments about the results provided by the system?

[« Back](#)

[Continue »](#)

Powered by [Google Docs](#)

[Report Abuse](#) - [Terms of Service](#) - [Additional Terms](#)

Evaluation of BiOSS system

Section 2 - Test case 3 (UK cancer registries)

(if you do not want to fill out this test case, click on the "Continue" button at the bottom of this page)

The UK is covered by 11 cancer registries, coordinated by the United Kingdom Association of Cancer Registries (UKACR). These registries, which have stored population-based information on cancer from more than 40 years ago, contain a common set of 24 epidemiological variables. In order to facilitate the future integration of these data with data from other registries, it is necessary find the best ontologies to semantically annotate these variables.

VARIABLES (24 terms): hospital, consultant, patient unit number, nhs number, forenames, surname, name at birth, address at time of diagnosis, postal code, sex, ethnic origin, date of birth, neoplasm location, morphology, laterality, stage, tumor grading, basis of diagnosis, date of diagnosis, treatment indicators, vital status, date of death, cause and place of death, post mortem.

1) Suppose that you have to choose ONE ontology (you can find a reference list of biomedical ontologies at <http://tinyurl.com/6dguy8u>) to solve this task. Please, provide a ranking (fill all positions as you can) of the ontologies you consider are the most adequate. You are allowed to use any website o tool you consider helpful.

Example: 1st: Gene Ontology; 2nd: NCI Thesaurus; 3rd: Medical Subject Headings, etc.

2) How much time did you take to answer the previous question? If you did not answer it, please, indicate why (e.g. I would need at least 1 hour, it is a too complex task, etc.)

3) Now, suppose that you are allowed to choose ONE OR SEVERAL biomedical ontologies to solve the task. Please, provide a ranking (fill all positions as you can) of the ontologies you consider are the most adequate. You are allowed to use any website o tool you consider helpful.

Example: 1st: Gene Ontology + NCI Thesaurus; 2nd: Gene Ontology + Medical Subject Headings + Galen Ontology; 3rd: Foundational Model of Anatomy; etc.

4) How much time did you take to answer the previous question? If you did not answer it, please, indicate why (e.g. I would need at least 1 hour, it is a too complex task, etc.)

5) If you had more time, do you think that the rankings you have provided would be better? How much time do you think you would need?

6) Have you used any tool or website to answer the previous questions? Which one/s?

7) Please, go to the BiOSS website (<http://bioss.ontologyselection.com>), choose "Test case 3 (UK cancer registries)" as input and execute it (button "Go!") both in "Single Output" and "Combined Output" modes. Examine the top 10 results provided by the system. How helpful do you think that the results are to the proposed task?

	Not helpful	Few helpful	Helpful	Very helpful
Single output	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Combined output	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

8) After examining the results provided by the system, would you modify the rankings provided by you? How would you modify them?

9) Any other comments about the results provided by the system?

[« Back](#) [Continue »](#)

Powered by [Google Docs](#)

[Report Abuse](#) - [Terms of Service](#) - [Additional Terms](#)

Evaluation of BiOSS system

Section 2 - Test case 4 (Anatomy)

(if you do not want to fill out this test case, click on the "Continue" button at the bottom of this page)

DESCRIPTION: Suppose that it is necessary to find the best ontology or ontologies to annotate a set of terms randomly chosen from the domain of anatomy.

TERMS (80 terms): dimorphism, biliary system, distal, dorsal, ectoderm, electrolyte, sural nerve, synapse, calcaneal tendon, endocardium, endoderm, epithelium, gonadotropins, heterophilic, homeothermic, squamous, stratified epithelium, set of meninges, paraganglion, radial nerve, dopamine receptor, syncytium, telogen, skull, cavity of stomach, surface of bone of foot, cranial cavity, spinal canal, thoracic cavity, gingiva, abdominopelvic cavity, spectrin, pericardial cavity, gallbladder, spleen, kidneys, set of muscles of abdomen, basilar artery, hepatic vein, uterine tubes, sagittal plane, ventral, vertebral canal, agger nasi, allantois, alveus, anconeus, annulus, hamstrings, quadrigeminus, macula, mammilla, masseter, mediasinus, mesencephalon, modiolus, mylohyoid, pampiniform, paraesthesia, parotid, pedicle, pennate, periosteum, platysma, proprioceptive, psoas, arch of aorta, axon, lysosome, mechanoreceptor, tibial tuberosity, perimysium, musculature of head, rectus capitis anterior, popliteus, midcarpal joint, obturator canal, long bone, skin of elbow, umbilical vein, supination.

1) Suppose that you have to choose ONE ontology (you can find a reference list of biomedical ontologies at <http://tinyurl.com/6dguy8u>) to solve this task. Please, provide a ranking (fill all positions as you can) of the ontologies you consider are the most adequate. You are allowed to use any website o tool you consider helpful.

Example: 1st: Gene Ontology; 2nd: NCI Thesaurus; 3rd: Medical Subject Headings, etc.

2) How much time did you take to answer the previous question? If you did not answer it, please, indicate why (e.g. I would need at least 1 hour, it is a too complex task, etc.)

3) Now, suppose that you are allowed to choose ONE OR SEVERAL biomedical ontologies to solve the task. Please, provide a ranking (fill all positions as you can) of the ontologies you consider are the most adequate. You are allowed to use any website o tool you consider helpful.

Example: 1st: Gene Ontology + NCI Thesaurus; 2nd: Gene Ontology + Medical Subject Headings + Galen Ontology; 3rd: Foundational Model of Anatomy; etc.

4) How much time did you take to answer the previous question? If you did not answer it, please, indicate why (e.g. I would need at least 1 hour, it is a too complex task, etc.)

5) If you had more time, do you think that the rankings you have provided would be better? How much time do you think you would need?

6) Have you used any tool or website to answer the previous questions? Which one/s?

7) Please, go to the BiOSS website (<http://bioss.ontologyselection.com>), choose “Test case 4 (Anatomy)” as input and execute it (button “Go!”) both in “Single Output” and “Combined Output” modes. Examine the top 10 results provided by the system. How helpful do you think that the results are to the proposed task?

	Not helpful	Few helpful	Helpful	Very helpful
Single output	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Combined output	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

8) After examining the results provided by the system, would you modify the rankings provided by you? How would you modify them?

9) Any other comments about the results provided by the system?

[« Back](#) [Continue »](#)

Powered by [Google Docs](#)

[Report Abuse](#) - [Terms of Service](#) - [Additional Terms](#)

Evaluation of BiOSS system

Section 2 - Test case 5 (Test case 1 with example ontology)

IMPORTANT: To fill out this test case, it is recommended that you have completed the test case 1 (breast cancer) before.

If you do not want to fill out this test case, click on the "Continue" button at the bottom of this page.

DESCRIPTION: Suppose that you are in the situation described in the test case 1 (breast cancer), but having into account a new ontology, called "Example ontology". This ontology contains only 30 concepts, but these concepts cover the 100% of the 30 initial terms in the test case 1. However, this ontology has been created individually. You do not know if it contains mistakes or not. It is not a shared ontology, and it is not a relevant ontology in the biomedical community. This test case is about giving your opinion about the position that should have this ontology in the ranking.

- 1) Please, go to the BiOSS website (<http://bioss.ontologyselection.com>), choose "Test case 5 (TC1 + Example ontology)" as input and execute it (button "Go!") in "Single Output" mode. What do you think about the position of the "Example ontology" in the ranking obtained?

The position of the "Example ontology" in the ranking is NOT adequate. It should be ranked higher	<input type="radio"/>
The position of the "Example ontology" in the ranking is NOT adequate. It should be ranked lower	<input type="radio"/>
The position of the "Example ontology" in the ranking is adequate	<input type="radio"/>
I don't know if it is adequate or not	<input type="radio"/>

- 2) Please, explain your previous answer

[« Back](#) [Continue »](#)

Powered by [Google Docs](#)

[Report Abuse](#) - [Terms of Service](#) - [Additional Terms](#)

Evaluation of BiOSS system

* Required

Section 3 - General questions

1) In general, how helpful do you think is the system in selecting the most adequate ontology or set of ontologies for a given context? *

	Not helpful	Few helpful	Helpful	Very helpful
Single output	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Combined output	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

2) What do you think are the positive features of the system? *

3) What do you think are the weak features of the system? Could you provide any suggestions to improve it? *

4) Any other comments?

[« Back](#) [Submit](#)

Powered by [Google Docs](#)

[Report Abuse](#) - [Terms of Service](#) - [Additional Terms](#)

Anexo III. Interfaz del Servicio Web

A continuación, se muestra el fichero WSDL (*Web Services Description Language*) del servicio de selección de ontologías que se ha implementado. Este fichero proporciona una descripción en detalle de la interfaz pública del servicio Web y del formato de los mensajes intercambiados.

```
<?xml version="1.0" encoding="UTF-8"?>
<wsdl:definitions xmlns:wsdl="http://schemas.xmlsoap.org/wsdl/"
xmlns:ns1="http://org.apache.axis2/xsd"
xmlns:ns="http://service.evaluator.java.imedir.udc.es"
xmlns:wsaw="http://www.w3.org/2006/05/addressing/wsdl"
xmlns:http="http://schemas.xmlsoap.org/wsdl/http/"
xmlns:ax23="http://TO.ontology.ontologies.java.imedir.udc.es/xsd"
xmlns:ax21="http://service.evaluator.java.imedir.udc.es/xsd"
xmlns:xs="http://www.w3.org/2001/XMLSchema"
xmlns:mime="http://schemas.xmlsoap.org/wsdl/mime/"
xmlns:soap="http://schemas.xmlsoap.org/wsdl/soap/"
xmlns:soap12="http://schemas.xmlsoap.org/wsdl/soap12/"
targetNamespace="http://service.evaluator.java.imedir.udc.es">
    <wsdl:documentation>BioSSWSERVICE</wsdl:documentation>
    <wsdl:types>
        <xss:schema xmlns:ax24="http://TO.ontology.ontologies.java.imedir.udc.es/xsd"
xmlns:ax22="http://service.evaluator.java.imedir.udc.es/xsd"
attributeFormDefault="qualified" elementFormDefault="qualified"
targetNamespace="http://service.evaluator.java.imedir.udc.es">
            <xss:import namespace="http://service.evaluator.java.imedir.udc.es/xsd"/>
            <xss:import
namespace="http://TO.ontology.ontologies.java.imedir.udc.es/xsd"/>
            <xss:element name="selectOntologies">
                <xss:complexType>
                    <xss:sequence>
                        <xss:element minOccurs="0" name="calculateCombinedOutput"
type="xs:boolean"/>
                        <xss:element minOccurs="0"
name="maxNumberOfCombinedOntologies" type="xs:int"/>
                        <xss:element maxOccurs="unbounded" minOccurs="0"
name="expandedTermsSets" nillable="true" type="ns:ArrayOfString"/>
                        <xss:element maxOccurs="unbounded" minOccurs="0"
name="ontologyIds" type="xs:long"/>
                        <xss:element minOccurs="0" name="wcc" type="xs:float"/>
                        <xss:element minOccurs="0" name="wsr" type="xs:float"/>
                        <xss:element minOccurs="0" name="wp" type="xs:float"/>
                        <xss:element minOccurs="0" name="wcckins" type="xs:float"/>
                        <xss:element minOccurs="0" name="wccinf" type="xs:float"/>
                        <xss:element minOccurs="0" name="wccscal" type="xs:float"/>
                        <xss:element minOccurs="0" name="wwiki" type="xs:float"/>
                        <xss:element minOccurs="0" name="wtwitt" type="xs:float"/>
                        <xss:element minOccurs="0" name="wbiop" type="xs:float"/>
                        <xss:element minOccurs="0" name="wpubm" type="xs:float"/>
                        <xss:element minOccurs="0" name="minContextCoverage"
type="xs:float"/>
                    </xss:sequence>
                </xss:complexType>
            </xss:element>
        </xss:schema>
    </wsdl:types>

```

```

        </xs:complexType>
    </xs:element>
    <xs:complexType name="ArrayOfString">
        <xs:sequence>
            <xs:element maxOccurs="unbounded" minOccurs="0" name="array"
nillable="true" type="xs:string"/>
        </xs:sequence>
    </xs:complexType>
    <xs:element name="selectOntologiesResponse">
        <xs:complexType>
            <xs:sequence>
                <xs:element maxOccurs="unbounded" minOccurs="0"
name="return" nillable="true" type="ax22:OutputElementService"/>
            </xs:sequence>
        </xs:complexType>
    </xs:element>
    <xs:element name="getOntologiesResponse">
        <xs:complexType>
            <xs:sequence>
                <xs:element maxOccurs="unbounded" minOccurs="0"
name="return" nillable="true" type="ax24:OntologyTO"/>
            </xs:sequence>
        </xs:complexType>
    </xs:element>
    <xs:element name="getExpandedTermSets">
        <xs:complexType>
            <xs:sequence>
                <xs:element maxOccurs="unbounded" minOccurs="0"
name="initialTerms" nillable="true" type="xs:string"/>
            </xs:sequence>
        </xs:complexType>
    </xs:element>
    <xs:element name="getExpandedTermSetsResponse">
        <xs:complexType>
            <xs:sequence>
                <xs:element maxOccurs="unbounded" minOccurs="0"
name="return" nillable="true" type="ns:ArrayOfString"/>
            </xs:sequence>
        </xs:complexType>
    </xs:element>
</xs:schema>
<xs:schema attributeFormDefault="qualified" elementFormDefault="qualified"
targetNamespace="http://service.evaluator.java.imedir.udc.es/xsd">
    <xs:complexType name="OutputElementService">
        <xs:sequence>
            <xs:element minOccurs="0" name="aggregatedScore"
type="xs:float"/>
            <xs:element minOccurs="0" name="ccScore" type="xs:float"/>
            <xs:element maxOccurs="unbounded" minOccurs="0"
name="ontologyIds" nillable="true" type="xs:long"/>
            <xs:element maxOccurs="unbounded" minOccurs="0"
name="ontologyNames" nillable="true" type="xs:string"/>
            <xs:element minOccurs="0" name="pScore" type="xs:float"/>
            <xs:element minOccurs="0" name="srScore" type="xs:float"/>
            <xs:element maxOccurs="unbounded" minOccurs="0"
name="termsMatched" nillable="true" type="xs:string"/>
            <xs:element maxOccurs="unbounded" minOccurs="0"
name="termsSources" nillable="true" type="xs:long"/>
        </xs:sequence>
    </xs:complexType>
</xs:schema>

```

```

<xs:schema attributeFormDefault="qualified" elementFormDefault="qualified"
targetNamespace="http://TO.ontology.ontologies.java.imedir.udc.es/xsd">
    <xs:complexType name="OntologyTO">
        <xs:sequence>
            <xs:element minOccurs="0" name="bioportalRefs" type="xs:int"/>
            <xs:element minOccurs="0" name="ontologyAlternativeNames"
nillable="true" type="xs:string"/>
                <xs:element minOccurs="0" name="ontologyDownloadLocation"
nillable="true" type="xs:string"/>
                    <xs:element minOccurs="0" name="ontologyId" type="xs:long"/>
                    <xs:element minOccurs="0" name="ontologyIdInSource"
nillable="true" type="xs:string"/>
                        <xs:element minOccurs="0" name="ontologyName" nillable="true"
type="xs:string"/>
                            <xs:element minOccurs="0" name="ontologySource" nillable="true"
type="xs:string"/>
                                <xs:element minOccurs="0" name="ontologyStoragePlace"
nillable="true" type="xs:string"/>
                                    <xs:element minOccurs="0" name="ontologyVersion" nillable="true"
type="xs:string"/>
                                        <xs:element minOccurs="0" name="pubmedRefs" type="xs:int"/>
                                        <xs:element minOccurs="0" name="twitterRefs" type="xs:int"/>
                                        <xs:element minOccurs="0" name="wikipediaRefs" type="xs:int"/>
                                    </xs:sequence>
                                </xs:complexType>
                            </xs:schema>
                        </wsdl:types>
                        <wsdl:message name="getOntologiesRequest"/>
                        <wsdl:message name="getOntologiesResponse">
                            <wsdl:part name="parameters" element="ns:getOntologiesResponse"/>
                        </wsdl:message>
                        <wsdl:message name="selectOntologiesRequest">
                            <wsdl:part name="parameters" element="ns:selectOntologies"/>
                        </wsdl:message>
                        <wsdl:message name="selectOntologiesResponse">
                            <wsdl:part name="parameters" element="ns:selectOntologiesResponse"/>
                        </wsdl:message>
                        <wsdl:message name="getExpandedTermSetsRequest">
                            <wsdl:part name="parameters" element="ns:getExpandedTermSets"/>
                        </wsdl:message>
                        <wsdl:message name="getExpandedTermSetsResponse">
                            <wsdl:part name="parameters" element="ns:getExpandedTermSetsResponse"/>
                        </wsdl:message>
                    <wsdl:portType name="BioSSWServicePortType">
                        <wsdl:operation name="getOntologies">
                            <wsdl:input message="ns:getOntologiesRequest"
wsaw:Action="urn:getOntologies"/>
                            <wsdl:output message="ns:getOntologiesResponse"
wsaw:Action="urn:getOntologiesResponse"/>
                        </wsdl:operation>
                        <wsdl:operation name="selectOntologies">
                            <wsdl:input message="ns:selectOntologiesRequest"
wsaw:Action="urn:selectOntologies"/>
                            <wsdl:output message="ns:selectOntologiesResponse"
wsaw:Action="urn:selectOntologiesResponse"/>
                        </wsdl:operation>
                        <wsdl:operation name="getExpandedTermSets">
                            <wsdl:input message="ns:getExpandedTermSetsRequest"
wsaw:Action="urn:getExpandedTermSets"/>
                            <wsdl:output message="ns:getExpandedTermSetsResponse"
wsaw:Action="urn:getExpandedTermSetsResponse"/>
                        </wsdl:operation>
                    </wsdl:portType>
                </wsdl:service>
            </wsdl:definitions>
        
```

```
</wsdl:operation>
</wsdl:portType>
<wsdl:binding name="BioSSWSERVICESoap11Binding" type="ns:BioSSWSERVICEPortType">
    <soap:binding transport="http://schemas.xmlsoap.org/soap/http"
style="document"/>
        <wsdl:operation name="getOntologies">
            <soap:operation soapAction="urn:getOntologies" style="document"/>
            <wsdl:input>
                <soap:body use="literal"/>
            </wsdl:input>
            <wsdl:output>
                <soap:body use="literal"/>
            </wsdl:output>
        </wsdl:operation>
        <wsdl:operation name="selectOntologies">
            <soap:operation soapAction="urn:selectOntologies" style="document"/>
            <wsdl:input>
                <soap:body use="literal"/>
            </wsdl:input>
            <wsdl:output>
                <soap:body use="literal"/>
            </wsdl:output>
        </wsdl:operation>
        <wsdl:operation name="getExpandedTermSets">
            <soap:operation soapAction="urn:getExpandedTermSets" style="document"/>
            <wsdl:input>
                <soap:body use="literal"/>
            </wsdl:input>
            <wsdl:output>
                <soap:body use="literal"/>
            </wsdl:output>
        </wsdl:operation>
    </wsdl:binding>
    <wsdl:binding name="BioSSWSERVICESoap12Binding" type="ns:BioSSWSERVICEPortType">
        <soap12:binding transport="http://schemas.xmlsoap.org/soap/http"
style="document"/>
            <wsdl:operation name="getOntologies">
                <soap12:operation soapAction="urn:getOntologies" style="document"/>
                <wsdl:input>
                    <soap12:body use="literal"/>
                </wsdl:input>
                <wsdl:output>
                    <soap12:body use="literal"/>
                </wsdl:output>
            </wsdl:operation>
            <wsdl:operation name="selectOntologies">
                <soap12:operation soapAction="urn:selectOntologies" style="document"/>
                <wsdl:input>
                    <soap12:body use="literal"/>
                </wsdl:input>
                <wsdl:output>
                    <soap12:body use="literal"/>
                </wsdl:output>
            </wsdl:operation>
            <wsdl:operation name="getExpandedTermSets">
                <soap12:operation soapAction="urn:getExpandedTermSets"
style="document"/>
                <wsdl:input>
                    <soap12:body use="literal"/>
                </wsdl:input>
                <wsdl:output>

```

```
        <soap12:body use="literal"/>
    </wsdl:output>
</wsdl:operation>
</wsdl:binding>
<wsdl:binding name="BioSSWSERVICEHttpBinding" type="ns:BioSSWSERVICEPortType">
    <http:binding verb="POST"/>
    <wsdl:operation name="getOntologies">
        <http:operation location="BioSSWSERVICE/getOntologies"/>
        <wsdl:input>
            <mime:content type="text/xml" part="getOntologies"/>
        </wsdl:input>
        <wsdl:output>
            <mime:content type="text/xml" part="getOntologies"/>
        </wsdl:output>
    </wsdl:operation>
    <wsdl:operation name="selectOntologies">
        <http:operation location="BioSSWSERVICE/selectOntologies"/>
        <wsdl:input>
            <mime:content type="text/xml" part="selectOntologies"/>
        </wsdl:input>
        <wsdl:output>
            <mime:content type="text/xml" part="selectOntologies"/>
        </wsdl:output>
    </wsdl:operation>
    <wsdl:operation name="getExpandedTermSets">
        <http:operation location="BioSSWSERVICE/getExpandedTermSets"/>
        <wsdl:input>
            <mime:content type="text/xml" part="getExpandedTermSets"/>
        </wsdl:input>
        <wsdl:output>
            <mime:content type="text/xml" part="getExpandedTermSets"/>
        </wsdl:output>
    </wsdl:operation>
</wsdl:binding>
<wsdl:service name="BioSSWSERVICE">
    <wsdl:port name="BioSSWSERVICEHttpSoap11Endpoint"
binding="ns:BioSSWSERVICESoap11Binding">
        <soap:address
location="http://193.147.41.219:8080/axis2/services/BioSSWSERVICE.BioSSWSERVICEHttps
oap11Endpoint//"/>
    </wsdl:port>
    <wsdl:port name="BioSSWSERVICEHttpSoap12Endpoint"
binding="ns:BioSSWSERVICESoap12Binding">
        <soap12:address
location="http://193.147.41.219:8080/axis2/services/BioSSWSERVICE.BioSSWSERVICEHttps
oap12Endpoint//"/>
    </wsdl:port>
    <wsdl:port name="BioSSWSERVICEHttpEndpoint"
binding="ns:BioSSWSERVICEHttpBinding">
        <http:address
location="http://193.147.41.219:8080/axis2/services/BioSSWSERVICE.BioSSWSERVICEHttpE
ndpoint//"/>
    </wsdl:port>
</wsdl:service>
</wsdl:definitions>
```


Referencias

- Alani, H., Brewster, C., & Shadbolt, N. (2006). *Ranking Ontologies with AKTiveRank*. International Semantic Web Conference 2006 (ISWC2006), Athens, GA, EE.UU.
- Alani, H., Noy, N., Shah, N., Shadbolt, N., & Musen, M. (2007). *Searching Ontologies Based on Content: Experiments in the Biomedical Domain*. 4th International Conference on Knowledge Capture (K-Cap), Whistler, BC, Canadá.
- Alexander, J., Freiling, M., Shulman, S., Staley, J., Rehfuss, S., & Messick, S. (1986). *Knowledge Level Engineering: Ontological Analysis*. 5th National Conference on Artificial Intelligence, Philadelphia, PA, EE.UU.
- Aristóteles. (350 a. C.). *Metaphysics*. Traducido por W. D. Ross. Londres: Oxford University Press, 1996.
- Arpírez, J., Gómez-Pérez, A., Lozano-Tello, A., & Pinto, H. (2000). Reference Ontology and (ONTO)² Agent: The Ontology Yellow Pages. *Knowledge and Information Systems*, 2(4), 387-412.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene Ontology: Tool for the Unification of Biology. *Nature Genetics*, 25, 25-29.
- Barta, R., Feilmayr, C., Pröll, B., Grün, C., & Werthner, H. (2009). *Covering the Semantic Space of Tourism: an Approach Based on Modularized Ontologies*. 1st Workshop on Context, Information and Ontologies (CIAO'09), Heraklion, Grecia.
- Bechhofer, S., Carr, L., Goble, C., Kampa, S., & Miles-Board, T. (2002). The Semantics of Semantic Annotation. *Lecture Notes In Computer Science*, 2519, 1152-1167.
- Berners-Lee, T. (1998). Semantic Web Road Map. Último acceso: Septiembre, 2010. Disponible en: <http://www.w3.org/DesignIssues/Semantic.html>.
- Berners-Lee, T. (1999). *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor*. San Francisco: Harper.

- Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, 284(5), 34-43.
- Blake, J. (2004). Bio-ontologies - Fast and Furious. *Nature Biotechnology*, 22(6), 773-774.
- Bodenreider, O. (2004). The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic Acids Research*, 32, 267-270.
- Bodenreider, O., & Stevens, R. (2006). Bio-ontologies: Current Trends and Future Directions. *Briefings in Bioinformatics*, 7(3), 256-274.
- Borgo, S., & Lesmo, L. (2008). *Formal Ontologies Meet Industry*: IOS Press.
- Borst, W. N. (1997). *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*. Universidad de Twente, Enschede, Holanda.
- Brank, J., Grobelnik, M., & Mladenic, D. (2005). *A Survey of Ontology Evaluation Techniques*. Conference on Data Mining and Data Warehouses (SiKDD 2005), Ljubljana, Eslovenia.
- Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., & Yergeau, F. (2000). Extensible Markup Language (XML) 1.0. W3C Recommendation. Último acceso: Abril, 2010. Disponible en: <http://www.w3.org/TR/xml/>.
- Brewster, C., Alani, H., Dasmahapatra, S., & Wilks, Y. (2004). *Data Driven Ontology Evaluation*. International Conference on Language Resources and Evaluation, Lisboa, Portugal.
- Brickley, D., & Guha, R. V. (2003). RDF Vocabulary Description Language 1.0: RDF Schema. *W3C Working Draft*, 23.
- Buggenhout, C. V., & Ceusters, W. (2005). A Novel View on Information Content of Concepts in a Large Ontology and a View on the Structure and the Quality of the Ontology. *International Journal of Medical Informatics*, 74(2-4), 125-132.
- Buitelaar, P., Eigner, T., & Declerck, T. (2004). *OntoSelect: A Dynamic Ontology Library with Support for Ontology Selection*. Sesión de Demostración en la International Semantic Web Conference, Hiroshima, Japón.
- Cantador, I., Fernández, M., & Castells, P. (2007). *Improving Ontology Recommendation and Reuse in WebCORE by Collaborative Assessments*. Workshop on Social and

- Collaborative Construction of Structured Knowledge. 16th International World Wide Web Conference (WWW 2007), Banff, Canadá.
- Castells, P. (2003). La Web Semántica. En C. Bravo & M. A. Redondo (Eds.), *Sistemas Interactivos y Colaborativos en la Web* (pp. 195–212): Ediciones de la Universidad de Castilla.
- CC:DA. (1999). *Committee on Cataloging: Description and Access (CC:DA), Task Force on Metadata and the Cataloging Rules. Informe Final.*
- Corcho, O. (2006). Ontology Based Document Annotation: Trends and Open Research Problems. *International Journal of Metadata, Semantics and Ontologies*, 1(1), 47-57.
- Côté, R., Jones, P., Apweiler, R., & Hermjakob, H. (2006). The Ontology Lookup Service, a Lightweight Cross-Platform Tool for Controlled Vocabulary Queries. *BMC Bioinformatics*, 7(1), 97.
- d'Aquin, M., Baldassarre, C., Gridinoc, L., Angeletou, S., Sabou, M., & Motta, E. (2007). *Watson: A Gateway for Next Generation Semantic Web Applications*. 6th International Semantic Web Conference, ISWC'07.
- Daconta, M. C., Obrst, L. J., & Smith, K. T. (2003). *The Semantic Web: a Guide to the Future of XML, Web Services, and Knowledge Management*: Wiley & Sons.
- Daniel, W., & Wayne, W. (2009). *Biostatistics: a Foundation for Analysis in the Health Sciences* (9th ed.). Nueva York: John Wiley & Sons.
- Davies, J., Fensel, D., & Van Harmelen, F. (2003). *Towards the Semantic Web. Ontology-driven Knowledge Management*: Wiley & Sons.
- Del Moral, A., Pazos, J., Rodríguez, E., Rodriguez-Paton, A., & Suarez, S. (2007). *Gestión del Conocimiento*. Madrid: Paraninfo.
- Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., et al. (2003). *SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation*. 12th International World Wide Web Conference, Budapest, Hungría.

- Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R., Peng, Y., et al. (2004). *Swoogle: a Search and Metadata Engine for the Semantic Web*. 13th ACM Conference on Information and Knowledge Management, Washington, WA, EE.UU.
- Ding, Y. (2005). Study of Design Issues on an Automated Semantic Annotation System. *AIS SIGSEMIS Bulletin*, 2, 45-51.
- Erdmann, M., Maedche, A., Schnurr, H., & Staab, S. (2000). *From Manual to Semi-Automatic Semantic Annotation: About Ontology-Based Text Annotation Tools*. COLING 2000 Workshop on Semantic Annotation and Intelligent Content, Luxemburgo.
- Euzenat, J., & Shvaiko, P. (2007). *Ontology Matching*. Nueva York: Springer-Verlag.
- Fensel, D., Horrocks, I., van Harmelen, F., Decker, S., Erdmann, M., & Klein, M. (2000). OIL in a Nutshell. *European Knowledge Acquisition Conference (EKAW-2000)*, 1–16.
- Fernández, M., Gómez-Pérez, A., Pazos, J., & Pazos, A. (1999). Building a Chemical Ontology using Methontology and the Ontology Design Environment. *IEEE Intelligent Systems and their Applications*, 14(1), 37-45.
- French, L., Lane, S., Law, T., Xu, L., & Pavlidis, P. (2009). Application and Evaluation of Automated Semantic Annotation of Gene Expression Experiments. *Bioinformatics*, 25(12), 1543.
- Gaeta, M., Orciuoli, F., & Ritrovato, P. (2009). Advanced Ontology Management System for Personalised e-Learning. *Knowledge-Based Systems*, 22(4), 292-301.
- Goclenius, R. (1613). *Lexicon Philosophicum*. Frankfurt.
- Gómez-Pérez, A. (1994). From Knowledge Based Systems to Knowledge Sharing Technology: Evaluation and Assessment. *Knowledge Systems Laboratory, Universidad de Stanford, CA, EE.UU.*
- Gómez-Pérez, A. (1995). *Some Ideas and Examples to Evaluate Ontologies*. 11th IEEE Conference on Artificial Intelligence Applications, Los Ángeles, CA, EE.UU.
- Gómez-Pérez, A. (1996). Towards a Framework to Verify Knowledge Sharing Technology. *Expert Systems with Applications*, 11(4), 519-529.

- Gómez-Pérez, A. (1999). *Evaluation of Taxonomic Knowledge in Ontologies and Knowledge Bases*. 12th Workshop on Knowledge Acquisition, Modeling and Management (KAW'99), Banff, Alberta, Canadá.
- Gómez-Pérez, A. (2004). Ontology Evaluation. En S. Staab & R. Studer (Eds.), *Handbook on Ontologies* (pp. 251–274). Berlín: Springer-Verlag.
- Gómez-Pérez, A., Fernández-López, M., & Corcho, O. (2004). *Ontological Engineering: With Examples from the Areas of Knowledge Management, E-Commerce and the Semantic Web*. Londres: Springer Verlag.
- Gómez-Pérez, A., Juristo, N., & Pazos, J. (1995). Evaluation and Assessment of Knowledge Sharing Technology. *Towards Very Large Knowledge Bases*, 289-296.
- Gómez-Pérez, A., & Rojas-Amaya, M. D. (1999). Ontological Reengineering for Reuse. *Knowledge Acquisition, Modeling and Management. Lecture Notes in Computer Science*, 1621/1999, 139-156.
- Grandi, F., Mandreoli, F., Martoglia, R., Ronchetti, E., Scalas, M., & Tiberio, P. (2009). Ontology-based Personalization of e-Government Services. En C. Mourlas & P. Germanakos (Eds.), *Intelligent User Interfaces* (pp. 167-203): IGI Global.
- Greenberg, J., Sutton, S., & Campbell, D. G. (2003). Metadata: A Fundamental Component of the Semantic Web. *Bulletin of the American Society for Information Science and Technology*, 29(4), 16-18.
- Gruber, T. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5(2), 199-220.
- Gruber, T. (1994). Introduction to the Bibliographic-Data Ontology. Último acceso: Julio, 2010. Disponible en: <http://www-ksl.stanford.edu/knowledge-sharing/ontologies/html/bibliographic-data.text.html>.
- Gruber, T. (2008). Collective Knowledge Systems: Where the Social Web Meets the Semantic Web. *Journal of Web Semantics*, 6(1), 4-13.
- Gruber, T., & Olsen, G. (1994). *An Ontology for Engineering Mathematics*. 4th International Conference on Principles of Knowledge Representation and Reasoning, San Mateo, CA, EE.UU.

- Gruber, T. R. (1995). Toward Principles for the Design of Ontologies used for Knowledge Sharing. *International Journal of Human Computer Studies*, 43(5), 907-928.
- Gruninger, M., & Fox, M. (1995). *Methodology for the Design and Evaluation of Ontologies*. IJCAI95's Workshop on Basic Ontological Issues in Knowledge Sharing, Montreal, Canadá.
- Gruninger, M., & Fox, M. S. (1994). *The Role of Competency Questions in Enterprise Engineering*. IFIP Workshop on Benchmarking, Theory and Practice, Trondheim, Noruega.
- Guarino, N. (1998). *Formal Ontology and Information Systems*. Formal Ontology in Information Systems (FOIS 1998), Trento, Italia.
- Guarino, N., & Welty, C. (2002). Evaluating Ontological Decisions with OntoClean. *Communications of the ACM*, 45(2), 61-65.
- Guarino, N., & Welty, C. A. (2009). An Overview of OntoClean. En S. Staab & R. Studer (Eds.), *Handbook on Ontologies* (pp. 201-220). Berlín: Springer-Verlag.
- Haase, K. (2004). *Context for Semantic Metadata*. 12th annual ACM International Conference on Multimedia, Nueva York, NY, EE.UU.
- Handschrift, S., & Staab, S. (2003). *Annotation for the Semantic Web*. IOS Press.
- Hayes, P. (1985). The Second Naïve Physics Manifesto. En Hoobs & Moore (Eds.), (pp. 71-108): Morgan Kaufmann Publishers Inc.
- Hong, T. P., Chang, W. C., & Lin, J. H. (2005). *A Two-Phased Ontology Selection Approach for Semantic Web*. 9th International Conference on Knowledge-Based & Intelligent Information & Engineering Systems, KES'05, Melbourne, Australia.
- Horrocks, I., & van Harmelen, F. (2000). *Reference Description of the DAML+ OIL Ontology Markup Language*. Defense Advanced Research Projects Agency (DARPA).
- Hovy, E. (2002). Comparing Sets of Semantic Relations in Ontologies. En R. Green, C. A. Bean & S. H. Myaeng (Eds.), *Semantics of Relationships: An Interdisciplinary Perspective* (pp. 91–110): Kluwer.

- Jiang, L., Zhan, J., Li, L., Shi, C., & An, N. (2008). *Utilizing User Behaviors with Semantic Metadata*. 5th International Conference on Information Technology: New Generations, Las Vegas, NV, EE.UU.
- Jones, M., & Alani, H. (2006). *Content-based Ontology Ranking*. 9th International Protégé Conference, Stanford, CA, EE.UU.
- Jonquet, C., Musen, M. A., & Shah, N. (2008). *A System for Ontology-Based Annotation of Biomedical Data*. International Workshop on Data Integration in The Life Sciences 2008, DILS'08, Evry, Francia.
- Jonquet, C., Musen, M. A., & Shah, N. H. (2010). Building a Biomedical Ontology Recommender Web Service. *Journal of Biomedical Semantics*, 1(S1), 1-18.
- Jonquet, C., Shah, N. H., & Musen, M. A. (2009). *Prototyping a Biomedical Ontology Recommender Service*. Bio-Ontologies 2009 Conference, Estocolmo, Suecia.
- Karmacharya, A., Cruz, C., Boochs, F., & Marzani, F. (2009). *ArcheoKM: Toward a Better Archaeological Spatial Datasets Management*. Computer Applications and Quantitative Methods in Archaeology (CAA), Williamsburg, Virginia, EE.UU.
- Karp, P. D., Chaudhri, V. K., & Thomere, J. (1999). XOL: An XML-based Ontology Exchange Language. *Versión 0.3*.
- Khelif, K., Dieng-Kuntz, R., & Barbry, P. (2007). An Ontology-based Approach to Support Text Mining and Information Retrieval in the Biological Domain. *Journal of Universal Computer Science*, 13(12), 1881-1907.
- Kiryakov, A., Popov, B., Terziev, I., Manov, D., & Ognyanoff, D. (2004). Semantic Annotation, Indexing, and Retrieval. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2(1), 49-79.
- Köhler, J., Munn, K., Rüegg, A., Skusa, A., & Smith, B. (2006). Quality Control for Terms and Definitions in Ontologies and Taxonomies. *BMC Bioinformatics*, 7(1), 212-224.
- Kudelka, M., Snasel, V., Lehecka, O., & El-Qawasmeh, E. (2006). Semantic Analysis of Web Pages Using Web Patterns. *Web Intelligence*, 329-333.

- Lacy, L. W. (2005). *Owl: Representing Information Using The Web Ontology Language*. Trafford Publishing.
- Lassila, O., & Hendler, J. (2007). Embracing Web 3.0. *IEEE Internet Computing*, 90-93.
- Lassila, O., & McGuinness, D. (2001). The Role of Frame-based Representation on the Semantic Web. *Linköping Electronic Articles in Computer and Information Science*, 6(5), 2001.
- Lassila, O., & Swick, R. (1999). Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendation. 22 de Febrero de 2009. Último acceso: Noviembre, 2010. Disponible en: <http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/>.
- Lewen, H., Supekar, K., Noy, N., & Musen, M. (2006). *Topic-Specific Trust and Open Rating Systems: An Approach for Ontology Evaluation*. 4th International EON Workshop Evaluating Ontologies for the Web, Edimburgo, Reino Unido.
- Liu, H., Hussain, F., Tan, C., & Dash, M. (2002). Discretization: An Enabling Technique. *Data Mining and Knowledge Discovery*, 6(4), 393-423.
- Liu, W., Jin, F., & Zhang, X. (2009). *Ontology-based User Modeling for e-Commerce System*. 3rd International Conference on Pervasive Computing and Applications (ICPCA 2008), Munich, Alemania.
- Lorhard, J. (1613). *Theatrum Philosophicum*. Basilea.
- Lozano-Tello, A., & Gómez-Pérez, A. (2004). ONTOMETRIC: A Method to Choose the Appropriate Ontology. *Journal of Database Management*, 15(2), 1-18.
- Luke, S., & Heflin, J. (2000). SHOE 1.01. Proposed Specification. *SHOE Project*.
- Macario, C., & Medeiros, C. (2009). A Framework for Semantic Annotation of Geospatial Data for Agriculture. *International Journal of Metadata, Semantics and Ontologies*, 4(1), 118-132.
- Maedche, A., & Staab, S. (2002). *Measuring Similarity between Ontologies*. 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW), Sigüenza, España.

- Maiga, G. (2009). *A Flexible Biomedical Ontology Selection Tool*. 5th International Conference on Computing and ICT Research (ICCIR'09), Kampala, Uganda.
- Marshall, C. C. (1998). *Toward an Ecology of Hypertext Annotation*. 9th ACM Conference on Hypertext and Hypermedia, Nueva York, EE.UU.
- Martínez-Romero, M., Vázquez-Naya, J., Rabuñal, J., Pita-Fernández, S., Macenlle, R., Castro-Alvariño, J., et al. (2010). Ontology and Complex Network of the Colorectal Cancer Drug Metabolism. *Current Drug Metabolism*, 11, 347-368.
- McCarthy, J. (1980). Circumscription - A Form of Non-monotonic Reasoning. *Artificial Intelligence*, 5(13), 27-39.
- McGuinness, D. L., & Van Harmelen, F. (2004). OWL Web Ontology Language Overview. W3C Recommendation. Último acceso: Noviembre, 2010. Disponible en: <http://www.w3.org/TR/owl-features/>.
- Meinong, A. (1904). The Theory of Objects. *Realism and the Background of Phenomenology (1960)*, 76–117.
- Min, H., Choi, J. Y., De Neve, W., Ro, Y. M., & Plataniotis, K. N. (2009). *Semantic Annotation of Personal Video Content Using an Image Folksonomy*. 16th IEEE International Conference on Image Processing (ICIP 2009), El Cairo, Egipto.
- Möller, M., Regel, S., & Sintek, M. (2009). *Radsem: Semantic Annotation and Retrieval for Medical Images*. 6th Annual European Semantic Web Conference (ESWC2009), Heraklion, Grecia.
- Moreira, D. A., & Musen, M. A. (2007). OBO to OWL: a Protégé OWL Tab to Read/Save OBO Ontologies. *Bioinformatics*, 23(14), 1868.
- National Center for Biomedical Ontology. (2010). BioPortal. Último acceso: Julio, 2010. Disponible en: <http://bioportal.bioontology.org>.
- Neches, R., Fikes, R., Finin, T., Gruber, T., & Patil, R. (1991). Enabling Technology for Knowledge Sharing. *AI Magazine*, 12(3), 36-56.
- Netzer, Y., Gabay, D., Adler, M., Goldberg, Y., & Elhadad, M. (2009). Ontology Evaluation Through Text Classification. *Advances in Web and Network Technologies, and Information Management* (pp. 210-221). Berlín: Springer-Verlag.

- Noy, N. F., Dorf, M., Griffith, N., Nyulas, C., & Musen, M. A. (2009). *Harnessing the Power of the Community in a Library of Biomedical Ontologies*. Workshop on Semantic Web Applications in Scientific Discourse, SWASD'09, Washington, WA, EE.UU.
- Noy, N. F., Guha, R., & Musen, M. A. (2005). *User Ratings of Ontologies: Who will Rate the Raters?* AAAI 2005 Spring Symposium on Knowledge Collection from Volunteer Contributions, Stanford, CA, EE.UU.
- Noy, N. F., & Hafner, C. D. (1997). The State of the Art in Ontology Design: A Survey and Comparative Review. *AI Magazine*, 18(3), 53.
- O'Reilly, T. (2005). What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. Último acceso: Noviembre, 2010. Disponible en: <http://oreilly.com/web2/archive/what-is-web-20.html>.
- Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). *The Pagerank Citation Ranking: Bringing Order to the Web*. Stanford, CA, EE.UU.: Informe Técnico. Stanford Digital Library Technologies Project, 1998.
- Pan, J. Z., Thomas, E., & Sleeman, D. (2006). *Ontosearch2: Searching and Querying Web Ontologies*. IADIS International Conference WWW/Internet Timisoara, Rumanía.
- Patel, C., Supekar, K., Lee, Y., & Park, E. K. (2003). *OntoKhoj: a Semantic Web Portal for Ontology Searching, Ranking and Classification*. 5th ACM International Workshop on Web Information and Data Management, Nueva Orleans, LA, EE.UU.
- Porzel, R., & Malaka, R. (2004). *A Task-based Approach for Ontology Evaluation*. ECAI Workshop on Ontology Learning and Population, Valencia, España.
- Quine, W. V. (1968). *Ontological Relativity and Other Essays*. New York: Columbia University Press.
- Raggett, D., Le Hors, A., & Jacobs, I. (1999). HTML 4.01 Specification. W3C Recommendation.
- Real Academia Española. (2001). *Diccionario de la Lengua Española (22^a edición)*. Madrid.

- Rindflesch, T. C., & Aronson, A. R. (1994). *Ambiguity Resolution while Mapping Free Text to the UMLS Metathesaurus*. 18th Annual Symposium on Computer Applications in Medical Care.
- Rotondo, F. (2010). *Geographical Information Systems and Ontologies: Two Instruments for Building Spatial Analysis Systems*. 2nd KES Symposium on Advances in Intelligent Decision Technologies (IDT 2010).
- Rubin, D., Shah, N., & Noy, N. (2007). Biomedical Ontologies: a Functional Perspective. *Briefings in Bioinformatics*, 9(1), 75-90.
- Saaty, T. L. (1977). A Scaling Method for Priorities in Hierarchical Structures. *Journal of Mathematical Psychology*, 15(3), 234-281.
- Sabou, M., Lopez, V., & Motta, E. (2006a). Ontology Selection for the Real Semantic Web: How to Cover the Queen's Birthday Dinner? *Managing Knowledge in a World of Networks. Lecture Notes in Artificial Intelligence*, 4248, 96-111.
- Sabou, M., Lopez, V., Motta, E., & Uren, V. (2006b). *Ontology Selection: Ontology Evaluation on the Real Semantic Web*. Evaluation of Ontologies on the Web Workshop, Edimburgo, Escocia.
- Scerri, S., Abela, C., & Montebello, M. (2005). *SemantExplorer: A Semantic Web Browser*. IADIS International Conference WWW/Internet, Lisboa, Portugal.
- Shah, N. H., & Musen, M. A. (2007). *Which Annotation did you Mean?* (Informe Técnico): Stanford Medical Informatics.
- Sheth, A., Bertram, C., Avant, D., Hammond, B., Kochut, K., & Warke, Y. (2002). Managing Semantic Content for the Web. *Internet Computing, IEEE*, 6(4), 80-87.
- Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., et al. (2007). The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration. *Nature Biotechnology*, 25(11), 1251-1255.
- Smith, B., & Brochhausen, M. (2008). Establishing and Harmonizing Ontologies in an Interdisciplinary Health Care and Clinical Research Environment. *Studies in Health Technology and Informatics*, 134, 219.

- Smith, B., & Welty, C. (2001). *Ontology: Towards a New Synthesis*. International Conference on Formal Ontology in Information Systems (FOIS2001), Maine, EE.UU.
- Sowa, J. F. (1984). *Conceptual Structures: Information Processing in Mind and Machine*. Reading, MA, EE.UU.: Addison-Wesley.
- Sowa, J. F. (2000). *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Belmont, CA, EE.UU.: Brooks/Cole.
- Staab, S., Maedche, A., & Handschuh, S. (2001). *An Annotation Framework for the Semantic Web*. 1st International Workshop on MultiMedia Annotation, Tokyo, Japón.
- Stevenson, M., & Wilks, Y. (2003). Word Sense Disambiguation. *The Oxford Handbook of Computational Linguistics* (pp. 249–265).
- Studer, R., Benjamins, V. R., & Fensel, D. (1998). Knowledge Engineering: Principles and Methods. *IEEE Transactions on Data & Knowledge Engineering*, 25(1-2), 161-197.
- Supekar, K. (2005). *A Peer-review Approach for Ontology Evaluation*. 8th International Protégé Conference, Madrid, España.
- Supekar, K., Patel, C., & Lee, Y. (2004). *Characterizing Quality of Knowledge on Semantic Web*. 17th International FLAIRS Conference, Miami, FL, EE.UU.
- Tan, H., & Lambrix, P. (2009). *Selecting an Ontology for Biomedical Text Mining*. Human Language Technology Conference, BioNLP Workshop, Colorado, EE.UU.
- Tartir, S., & Arpinar, I. B. (2007). *Ontology Evaluation and Ranking using OntoQA*. 1st IEEE International Conference on Semantic Computing (ICSC'07), Irvine, CA, EE.UU.
- Ungrangsi, R., Anutariya, C., & Wuwongse, V. (2008). CombiSQORE: a Combinative Ontology Retrieval System for Next Generation Semantic Web Applications. *IEICE Transactions on Information and Systems*, 91(11), 2616.
- Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., et al. (2006). Semantic Annotation for Knowledge Management: Requirements and a Survey

- of the State of the Art. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4(1), 14-28.
- Uschold, M., & Grüninger, M. (1996). Ontologies: Principles, Methods and Applications. *Knowledge Engineering Review*, 11(2), 93-136.
- Vilches-Blázquez, L. M., Ramos, J. A., López-Pellicer, F. J., Corcho, O., & Nogueras-Iso, J. (2009). An Approach to Comparing Different Ontologies in the Context of Hydrographical Information. En S. B. Heidelberg (Ed.), *Information Fusion and Geographic Information Systems* (Vol. 4, pp. 193-207). Berlín: Springer.
- Völker, J., Vrande, D., Sure, Y., & Hotho, A. (2008). AEON - An Approach to the Automatic Evaluation of Ontologies. *Applied Ontology*, 3(1), 41-62.
- Wang, X., Guo, L., & Fang, J. (2008). *Automated Ontology Selection Based on Description Logic*. 12th International Conference on Computer Supported Cooperative Work in Design, CSCWD'08, Xi'an, China.
- Welty, C., & Ide, N. (1999). Using the Right Tools: Enhancing Retrieval from Marked-up Documents. *Computers and the Humanities*, 33(1), 59-84.
- Whetzel, P. L., Shah, N. H., Noy, N. F., Dai, B., Dorf, M., Griffith, N., et al. (2009). BioPortal: Ontologies and Integrated Data Resources at the Click of a Mouse. *Nucleic Acids Research*, 4, 1-4.
- Xu, H., Zhou, X., Wang, M., Xiang, Y., & Shi, B. (2009). *Exploring Flickr's Related Tags for Semantic Annotation of Web Images*. ACM International Conference on Image and Video Retrieval (CIVR '09), Isla de Santorini, Grecia.
- Yesilada, Y., Harper, S., Goble, C., & Stevens, R. (2003). *Ontology Based Semantic Annotation for Enhancing Mobility Support for Visually Impaired Web Users*. KCAP 2003 Workshop on Knowledge Markup and Semantic Annotation, Sanibel, FL, EE.UU.
- Zhang, Y., Vasconcelos, W., & Sleeman, D. (2004). *OntoSearch: An Ontology Search Engine*. 24th SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, Cambridge, Reino Unido.

Zhang, Z., Zhang, C., & Ong, S. (2000). *Building an Ontology for Financial Investment*. 2nd International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2000), Hong Kong, China.