UNIVERSIDAD DE ÁLCALA DE HENARES

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN, ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA

TESIS DOCTORAL

APLICACIONES DE LA EXPANSIÓN DE CONSULTAS BASADA EN ONTOLOGÍAS DE DOMINIO A LA BÚSQUEDA DE OBJETOS DE APRENDIZAJE EN REPOSITORIOS.

Autor:

Alejandra Andrea Segura Navarrete

Licenciado en las Cs. De la Computación e Informática

Director:

Dr. Salvador Sánchez Alonso

Doctor en Informática

Universidad de Alcalá de Henares

Co-director:

Dr. Manuel E. Prieto Méndez

Doctor en Ciencias

Universidad de Castilla – La Mancha

Alcalá de Henares, 2010

Gracias a Dios…

Índice General

RESUMEN **IX**

ABSTRACT **XI**

AGRADECIMIENTOS **XIII**

CAPÍTULO 1. INTRODUCCIÓN A LA INVESTIGACIÓN **1**

1.1. INTRODUCCIÓN **1**
1.2. ESTRUCTURA DEL DOCUMENTO **2**

CAPÍTULO 2. ESTADO DE LA CUESTIÓN **5**

2.1. CONCEPTO Y ESTADO ACTUAL DEL ÁREA DE OBJETOS DE APRENDIZAJE **5**
2.1.1. OBJETOS DE APRENDIZAJE 6
2.1.2. METADATOS Y ESTÁNDARES DE METADATOS 7
2.2. REPOSITORIOS DE OBJETOS DE APRENDIZAJE **10**
2.2.1. FUNCIONALIDADES DE LOS REPOSITORIOS 11
2.2.2. CARACTERIZACIÓN DE LOS REPOSITORIOS 12
2.2.3. INTEROPERABILIDAD ENTRE REPOSITORIOS DE OBJETOS DE APRENDIZAJE 14
2.2.4. MECANISMOS DE ACCESO A LOS OBJETOS DE APRENDIZAJE 15
2.2.4.1 Servicio Web de búsqueda en repositorios 16
2.2.4.2 Open knowledge initiative 16
2.2.4.3 Búsqueda /recuperación vía servicio Web y vía URL 17
2.2.4.4 Capa de comunicación EduSource 17
2.2.4.5 Protocolo para la recolección automática de metadatos 18
2.2.4.6 Simple query interface 20
2.2.5. LENGUAJES DE CONSULTAS EN REPOSITORIOS 22
2.2.6. FORMATO DE LOS RESULTADOS RECUPERADOS DEL REPOSITORIO 24
2.2.7. ANÁLISIS DE LAS FUNCIONALIDADES DE BÚSQUEDA EN REPOSITORIOS 24
2.3. REPRESENTACIÓN DEL CONOCIMIENTO **27**
2.3.1. ONTOLOGÍAS 28
2.3.1.1 Clasificaciones 29
2.3.1.2 Lenguajes ontológicos 31
2.3.2. TESAUROS 35
2.3.3. COMPARACIÓN ENTRE ONTOLOGÍAS Y TESAUROS 37
2.4. RECUPERACIÓN DE INFORMACIÓN **37**
2.4.1. MÉTRICAS DE DESEMPEÑO DE LOS SISTEMAS RECUPERACIÓN DE INFORMACIÓN 41
2.4.2. RANKING DE RELEVANCIA EN LOS SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN 42
2.4.2.1 Algoritmo de ranking PageRank 42
2.4.2.2 Algoritmo de ranking HITS 43
2.4.2.3 Algoritmo de ranking SALSA 44
2.4.2.4 Ranking personalizado 45

2.4.2.5 Ranking de relevancia en la recuperación de objetos de aprendizaje 46
2.4.3. ONTOLOGÍAS EN LA RECUPERACIÓN DE INFORMACIÓN 47
2.4.4. EXPANSIÓN DE CONSULTAS 48
2.4.5. EXPANSIÓN DE CONSULTAS BASADAS EN MODELOS DE CONOCIMIENTO 53
2.4.5.1 Revisión de propuestas para la expansión de consultas basadas en modelos de conocimiento dependientes de la colección 54
2.4.5.2 Revisión de propuestas para la expansión de consultas basadas en modelos de conocimiento independientes de la colección 55
2.5. CARACTERIZACIÓN DE LA INVESTIGACIÓN 62

CAPÍTULO 3. PLANTEAMIENTO DEL PROBLEMA 69

3.1. PROBLEMAS GENERADOS POR EL USO DE ONTOLOGÍAS EN LA EXPANSIÓN DE CONSULTAS 69
3.2. PROBLEMAS DETECTADOS EN LA BÚSQUEDA DE OBJETOS DE APRENDIZAJE EN REPOSITORIOS 70
3.3. PROBLEMAS EN LA EXPANSIÓN DE CONSULTAS BASADAS EN MODELOS DE CONOCIMIENTO APLICADAS EN REPOSITORIOS DE OBJETOS DE APRENDIZAJE 75
3.4. DELIMITACIÓN DEL ÁMBITO DEL PROBLEMA 77
3.5. HIPÓTESIS DE PARTIDA 79
3.6. OBJETIVOS 79
3.7. MÉTODO GENERAL 80

CAPÍTULO 4. DESCRIPCIÓN DE LA SOLUCIÓN 83

4.1. PLANTEAMIENTO DE LA SOLUCIÓN PROPUESTA 83
4.1.1. TIPOS DE RELACIONES 83
4.1.2. DISTANCIA SEMÁNTICA 84
4.1.3. TIPOS DE EXPANSIONES 86
4.1.4. SOLUCIÓN AL PROBLEMA DE LA CORRESPONDENCIA EN LA ONTOLOGÍA 88
4.1.5. SOLUCIÓN AL PROBLEMA DE LA CALIDAD DEL MODELO DE CONOCIMIENTO 89
4.1.6. SOLUCIÓN AL PROBLEMA EN LA GENERACIÓN DE LA CADENA DE BÚSQUEDA EXPANDIDA 89
4.1.7. SOLUCIÓN AL PROBLEMA ASOCIADO AL TRATAMIENTO DE RESULTADOS DUPLICADOS 90
4.1.8. RESUMEN DEL PLANTEAMIENTO 97
4.2. DISEÑO DE LA SOLUCIÓN 97
4.2.1. ARQUITECTURA TÉCNICA 98
4.2.2. DESCRIPCIÓN FUNCIONAL DEL PROTOTIPO DE IMPLEMENTACIÓN 100

CAPÍTULO 5. EVALUACIÓN 105

5.1. CONTEXTO 105
5.1.1. ONTOLOGÍA GENE 105
5.1.2. REPOSITORIO MERLOT 106
5.2. PROCESO 107
5.3. EXTRACCIÓN DE CONSULTAS DE PRUEBA 108

5.4. EXPANSIÓN 111
5.5. BÚSQUEDA EN EL REPOSITORIO 115
5.6. FILTRADO DE LOS RESULTADOS 122
5.7. EVALUACIÓN DE RELEVANCIA 124
5.8. CONTRASTE DE RESULTADOS 128
5.9. LIMITACIONES 132
5.10. RESULTADOS 133
5.11. DISCUSIÓN DE LOS RESULTADOS POR TIPO DE EXPANSIÓN 147
5.12. DISCUSIÓN DE LOS RESULTADOS INDEPENDIENTEMENTE DEL TIPO DE EXPANSIÓN 151

CAPÍTULO 6. CONCLUSIONES 157

6.1. VERIFICACIÓN DE LA HIPÓTESIS Y LOS OBJETIVOS 157
6.2. CONCLUSIONES 160
6.3. APORTACIONES 161
6.4. LÍNEAS DE TRABAJO FUTURAS 162
6.5. PUBLICACIONES Y CONTRIBUCIONES DERIVADAS 164
6.6. OTRAS PUBLICACIONES 164
6.6.1. REVISTAS O CAPÍTULOS DE LIBROS 165
6.6.2. CONGRESOS INTERNACIONALES 165
6.6.3. CONGRESOS NACIONALES 167

REFERENCIAS 169

ANEXOS 185

ANEXO A : RESUMEN DE RESULTADOS OBTENIDOS DESDE EL REPOSITORIO PARA EL EXPERIMENTO 187

ANEXO B : LISTA DE RESULTADOS PONDERADOS, CASO DE ESTUDIO 191

ANEXO C : INSTRUMENTO PARA LA EVALUACIÓN DE LA RELEVANCIA DE LOS RESULTADOS 193

ANEXO D : EXTRACTO DE LA IMPLEMENTACIÓN DEL PROTOTIPO 195

Índice Tablas

Tabla 1: Estándares de metadatos según el contexto de aplicación. ____ *8*
Tabla 2: Perfiles de aplicación y esquemas de metadatos educacionales basados en IEEE-, IMS y DCMES. ____ *9*
Tabla 3: Síntesis de la evaluación realizada a las opciones de búsquedas provistas en los repositorios. *25*
Tabla 4: Tipos de ontologías según distintos criterios de clasificación. ____ *30*
Tabla 5: Características relevantes de los principales lenguajes ontológicos. Síntesis desde (McGuinness & van Harmelen, 2004; Motik, Patel-Schneider, & Parsia, 2009; Samper Zapater, 2006). _*32*
Tabla 6: Comparativa Lenguajes Ontológicos. Análisis inicial obtenido desde (Gómez-Pérez et al., 2004) y actualizado por el autor. ____ *33*
Tabla 7: Parte de la especificación de términos y relaciones en el tesauro ETB-LRE MEC-CCAA. ____ *36*
Tabla 8 : Síntesis de las características de las propuestas revisadas con mayor relevancia para nuestra investigación. ____ *64*
Tabla 9: Diferencias en los vocabularios y espacios de valores utilizados en los repositorios para los campos nivel educativo y tipo de recurso. ____ *74*
Tabla 10: Síntesis de los aspectos diferenciadores de la invetigación. ____ *78*
Tabla 11: Ejemplo de expansión, relación is_a: hijos. ____ *89*
Tabla 12: Lista de OA recuperados del repositorio (MERLOT) para los conceptos expandidos de la consulta T3 y el tipo de expansión is_a: hijos. ____ *91*
Tabla 13: Resultados de la re-ponderación y eliminación de resultados repetidos (objetos de aprendizaje). ____ *92*
Tabla 14: Listas individuales de los 10 primeros OA recuperados por cada tipo de expansión aplicado a la consulta T3. ____ *93*
Tabla 15: Listas de OA únicas. Ambas listas ponderan mejor un tipo de expansión distinto. ____ *96*
Tabla 16: Colección de Syllabus utilizados en este caso de estudio. ____ *109*
Tabla 17: Lista de consultas de prueba. ____ *110*
Tabla 18: Resumen de la expansión de términos. ____ *114*
Tabla 19: Cantidad de conceptos expandidos por consulta y frecuencia en el total de expansiones realizadas. ____ *115*
Tabla 20: Ejemplo resultados encontrados y recuperados desde el repositorio, para la consulta T17 Translation y tipo de expansión de sinónimo exacto (syn_exa). ____ *117*
Tabla 21: Resumen del proceso de búsqueda. ____ *118*
Tabla 22: Análisis de frecuencia de la cantidad de repeticiones de cada OA recuperádo. ____ *119*
Tabla 23: Análisis de frecuencia de la cantidad de repeticiones de cada OA recuperado único para una misma consulta y tipo de expansión. ____ *119*
Tabla 24: Análisis de frecuencia de la cantidad de repeticiones de cada OA recuperado único para una misma consulta. ____ *119*
Tabla 25: Eliminación de duplicados en las listas de resultados obtenidos de los términos expandidos, consulta T15 meiosis, tipo expansión is_a: hermanos. ____ *120*
Tabla 26: Lista de OA recuperados, ejemplo de la consulta original OR expansión, Consulta T15 Meiosis. ____ *121*
Tabla 27: Resultados de la búsqueda en el repositorio MERLOT después de aplicadas las restricciones (R1∩R2) ∩R3. ____ *123*
Tabla 28: Niveles para la interpretación del coeficiente Kappa. ____ *126*
Tabla 29: Coeficiente Kappa. Resumen para los 3 niveles de respuesta (relevante, parcialmente y no relevante) con 3 evaluadores. ____ *126*
Tabla 30: Análisis no paramétrico de Rho Spearman. Asociación general entre los niveles de relevancia, otorgada por los tres expertos, a los OA recuperados para cada consulta. ____ *127*
Tabla 31: Resultados consulta Original, T3 DNA replication. ____ *134*

Tabla 32: Lista de resultados por tipo de expansión en la consulta T3 DNA replication. Expansión is_a: hermanos (isa_her) y is_a: hijos (isa_hij). 135
Tabla 33: Resultados Consulta Original, T6 chromosome. 136
Tabla 34: Lista de resultados por tipo de expansión en la consulta T6 chromosome. Expansión is_a: hermanos (isa_her) y is_a: padre (isa_exa). 137
Tabla 35: Resultados consulta original T15 meiosis Lista de resultados por tipo de expansión en la consulta T15 meiosis. Expansión is_a: hermanos (isa_her). 138
Tabla 36: Resultados Consulta Original, T16 transcription. 139
Tabla 37: Lista de resultados por tipo de expansión en la consulta T16 transcription. Expansión is_a: hermanos (isa_her). 140
Tabla 38: Lista de resultados por tipo de expansión en la consulta T16 transcription. Expansión part_of: el todo (par_exa) y part_of: las partes (par_otr). 141
Tabla 39: Resultados Consulta Original, T17 translation 142
Tabla 40: Lista de resultados por tipo de expansión en la consulta T17 translation. Expansión sinónimo exacto (syn_exa). 143
Tabla 41: Lista de resultados por tipo de expansión en la consulta T17 translation. Expansión is_a: hermanos (isa_her) y part_of: el todo (par_exa). 144
Tabla 42: Lista de resultados por tipo de expansión en la consulta T17 translation. Expansión part_of: las partes (par_otr) y regulado por (reg_por). 145
Tabla 43: Resumen de los resultados de las métricas de novedad, cobertura y precisión para cada consulta y tipo de expansión. 146
Tabla 44: Lista de OA recuperados por la consulta original T17: translation y las listas individuales para algunos tipos de expansiones. 148
Tabla 45: Novedad y Cobertura en las distintas listas de resultados. 152

Índice de Figuras

Figura 1: Intersección de los elementos de metadatos. Esquema reproducido desde (Xiang et al., 2003). 10
Figura 2: Repositorios según su contenido. 11
Figura 3: Distribución del uso de estándares de metadatos en los repositorios de OA. Tabla de datos obtenidos de (McGreal et al., 2008). 14
Figura 4: Sistema de repositorios federados, ejemplo de la búsqueda federada en Ariadne. 15
Figura 5: Comunidades soportadas en la infraestructura EduSource. Fuente (Hatala et al., 2004b) pp. 21. 18
Figura 6: Intercambio de metadatos en El estándar OAI-PMH. 19
Figura 7: Intercambio de datos en SQI. 20
Figura 8: Implementación SQI en Java. 21
Figura 9: Implementación SQI en SOAP. 21
Figura 10. Interfaz SQI tester. 21
Figura 11: Resultados de la evaluación de la forma de búsqueda utilizada en los repositorios. 26
Figura 12: Representación de una Tripleta RDF. 31
Figura 13: A la izquierda el fenómeno co-citación y a la derecha el fenómeno co-referencia con el algoritmo HITS. 44
Figura 14: Intersección de la red semántica para la palabra mountain (synset 1) y la palabra top (synset 3). Los nodos comunes en este ejemplo son location y hill (ilustración reproducida de (Navigli & Velardi, 2003), pp. 44.) 56
Figura 15: Parte de la ontología CSO. Ilustración reproducida desde (Zou et al., 2008), pp. 454. 59
Figura 16: Árbol de intención del usuario, según algoritmo Lee et al. (2008). 61
Figura 17: Efecto de la profundidad de la ontología en el algoritmo de expansión propuesto por Lee et. al. (2008). 66
Figura 18: Ejemplo de opciones para la formulación de consultas a través de la navegación en Tesauro ETB-LRE MEC-CCAA, en el repositorio AGREGA. http://www.proyectoagrega.es/default/Inicio. 71
Figura 19: Ejemplo de opciones de refinamiento manual de consultas en el repositorio Gateway. 72
Figura 20: Ejemplo de opciones de refinamiento manual de consultas en el repositorio Laclo. 72
Figura 21: Ejemplo de la formulación de consultas a través de la navegación en una ontología de dominio (OA-AE), en el repositorio Organic. 73
Figura 22: Recuperación de información Web versus búsqueda en repositorios de OA. 78
Figura 23: Ejemplo tipo de relaciones modeladas en la ontología de Turismo (ProtegeWiki, 2010). 84
Figura 24: Ejemplo de distancia semántica utilizando parte de la ontología Food (ProtegeWiki, 2010) 85
Figura 25: Parte de la ontología Gene donde se representan las relaciones del tipo is_a y part-of entre los conceptos representados. Imágenes capturadas mediante OBO-Edit software para la edición de ontologías. http://oboedit.org/. 87
Figura 26: Parte de la ontología Gene donde se representan las relaciones del tipo is_a y part_of entre los conceptos representados. 88
Figura 27: Esquema para la generación de las listas de resultados por tipo de expansión 93
Figura 28: Esquema para la generación de las listas únicas de expansión. 94
Figura 29: Arquitectura funcional que da soporte a la expansión de consultas basada en ontologías en el contexto de la búsqueda de OA en repositorios. 98
Figura 30: Arquitectura técnico-funcional del prototipo de implementación. 101
Figura 31: Proceso de evaluación 108
Figura 32: Parte de la Ontología Gene donde se ejemplifican algunas de las relaciones básicas y específicas de una ontología en el dominio de Genética. 112

Figura 33: Gráfica del resumen de los resultados de la expansión en las consultas que coinciden exactamente con algún concepto en la ontología y en las consultas donde se utilizó el concepto más cercano. _____ *114*
Figura 34: Gráfica de resumen de los resultados de las consultas ejecutadas en el repositorio. _____ *118*
Figura 35: Métricas de cobertura y novedad para una consulta de ejemplo tomada de (Baeza-Yates & Ribeiro-Neto, 1999), pp. 83. _____ *130*
Figura 36: Ejemplo de estimación de cobertura y novedad. Conjunto de resultados relevantes conocidos U. _____ *131*
Figura 37: Ejemplo de estimación de cobertura y novedad. Conjunto de resultados relevantes desconocidos. _____ *131*
Figura 38: Niveles de precisión de los resultados recuperados por la consulta T17, original y expansiones. _____ *147*
Figura 39: Novedad para cada consulta y tipo de expansión. _____ *149*
Figura 40: Cobertura para cada consulta y tipo de expansión. _____ *150*
Figura 41: Promedio de cobertura, novedad y precisión. _____ *151*
Figura 42: Resumen de la métrica de Novedad para las distintas listas de análisis. _____ *154*
Figura 43: Resumen de la métrica de Precisión para las distintas listas de análisis. _____ *155*
Figura 44: Arquitectura propuesta ampliada según ésta línea de trabajo futura _____ *163*

Resumen

En el campo del e-learning se realizan grandes esfuerzos dedicados al mejoramiento del proceso de enseñanza – aprendizaje, uno de ellos está dirigido a motivar el uso y reutilización de los recursos digitales en repositorios. En un comienzo, los esfuerzos se dirigieron a aumentar la cantidad de recursos disponibles, hoy en día existe una gran cantidad de recursos almacenados en repositorios heterogéneos, por lo tanto el desafío se traslada a mejorar y hacer más eficientes las formas de buscar, seleccionar, localizar y acceder a recursos dispersos y distribuidos en repositorios.

Dentro de esta línea, el objetivo de esta tesis es proponer una estrategia para la expansión de consultas basadas en ontologías de dominio que permita al diseñador instruccional obtener, desde un repositorio, objetos de aprendizaje relevantes para el diseño de sus cursos o la composición de otros recursos más complejos. Para lograr este objetivo se analizan las propuestas de expansión de consultas ya sea en el campo de la recuperación de información en general o específicamente, en los repositorios de objetos de aprendizaje. A partir de lo anterior, se establecen los criterios para la expansión de consultas basada en ontologías, se define la forma como serán abordados los problemas detectados, y por último, se formula, diseña e implementa la estrategia de expansión de consultas basada en ontología de dominio aplicada en el contexto de la búsqueda de objetos de aprendizaje en repositorios.

Para la evaluación de nuestra propuesta se diseña un experimento dentro del dominio de genética, utilizando la ontología Gene como base de conocimiento y el repositorio MERLOT como proveedor de los objetos de aprendizaje en este dominio. Las consultas de prueba se definen a partir de los contenidos tratados en un conjunto de cursos de genética publicados en la Web por instituciones de educación superior para el año 2009. La evaluación de la relevancia de los resultados es realizada por 3 expertos en el dominio. El análisis de la concordancia y asociación entre las evaluaciones de los expertos es realizado por medio del análisis de *Kappa* de *Cohen* y el coeficiente de correlación de *Spearman*. Finalmente, la efectividad de la propuesta de expansión se evalúa a partir de las métricas de cobertura y novedad aplicadas a los resultados recuperados de las consultas con y sin expansión.

La principal aportación de nuestra propuesta es una estrategia para la expansión de consultas basada en ontologías de dominio que permita al diseñador instruccional obtener resultados relevantes que sin la expansión no podrían ser recuperados desde los repositorios de objetos de aprendizaje. Suponemos que en la medida que los diseñadores intruccionales puedan acceder a recursos relevantes es posible contribuir en la calidad de los cursos e-learning o en la calidad de los nuevos recursos creados a partir de ellos. Cabe destacar que la efectividad de nuestra propuesta se ve afectada por el sistema de recuperación utilizado en cada repositorio, la calidad de los recursos almacenados y su etiquetado, así como la completitud y calidad de la base de conocimiento utilizada para la expansión.

Keywords: Repositorios de objetos de aprendizaje, ontologías, expansión de consultas.

Abstract

In the e-learning field, great efforts are made to improve the learning process. One of these is focused on motivating the use and the reuse of digital resources in repositories. Initially, efforts were focused on increasing the amount of available resources. Nowadays, there are a lot of resources stored in repositories so the challenge has become that of improving the search, selection and access to resources scattered and distributed throughout the repositories.

In this sense, the objective of this thesis is to propose a strategy for query expansion based on domain ontologies which allow the instructional designer to find objects in a repository which are relevant for the design of their courses or to compose other, more complex resources. To achieve this goal, we analyse the proposals of query expansion either in the field of information retrieval in general, or specifically on learning object repositories. Based on the aforementioned, the criteria for query expansion based on ontologies have been established, and the ways to resolve the problems identified have been defined. Finally we propose, design and implement the query expansion strategy based on ontology domain applied in the context of the search for learning objects in repositories.

For our proposal, an experiment in the area of genetics has been designed using the Gene Ontology as a knowledge base and the MERLOT repository as a provider of learning objects in this domain. The test queries are defined using the contents dealt with in a set of courses in genetics published on the Web by higher education institutions in 2009. The evaluation of the results is carried out by three experts in the field. The analysis of the correlation and association between the experts' evaluations is carried out through the analysis of Cohens' Kappa, and Spearman's correlation coefficient. Finally, the effectiveness of the query expansion proposal is evaluated based on the metrics of coverage and novelty applied to the retrieved results of queries, with and without expansion.

The main contribution to this research is a strategy for query expansion based on domain ontologies to allow the instructional designer to obtain relevant results that, without the expansion, could not be retrieved by the repositories of learning objects. We expect that if the instructional designer can access relevant resources, this may contribute to the quality of e-learning courses or the quality of the new resources created from them. It is worth noting that the effectiveness of our proposal is affected by the information retrieval system used in each repository, the quality of stored resources and their labelling, and the completeness and quality of the knowledge base used for expansion.

Keywords: learning object repository, ontology, query expansion.

Agradecimientos

Que lejano me parecía el día en que tendría que escribir estos agradecimientos, y ahora aquí estoy.

En lo personal quiero agradecer el apoyo de mi esposo, mi familia y mis amigos, sé que ellos han compartido mis logros y también han sufrido conmigo todas las vicisitudes de este proceso. Gonzalo, nunca lo hubiese logrado sin tenerte a mi lado, dejaste muchas cosas por estar conmigo, mis logros son también los tuyos.

Quiero agradecer a mis padres quienes han sido fundamentales cada día, espero aprender de ustedes y enseñar a mis hijos el valor de la perseverancia, del trabajo, la honestidad, el respeto y la humildad. No puedo dejar de emocionarme al intentar expresar en palabras lo que ha significado mi hermano César durante todo este tiempo, es mi amigo, consejero y defensor incansable, gracias por estar siempre a mi lado.

Tengo la bendición de contar con buenos amigos, a mis amigas Claudia M. y Claudia V., gracias por ser incondicionales aún en la distancia. A Christian quien es como un hermano, tu apoyo ha sido imprescindible y a Gustavo, amigo gracias por tu amistad y compañerismo, amigos sin ustedes esto hubiese sido imposible. A Elizabeth, gracias por confiar y ayudarme en todo.

En lo profesional quiero agradecer a los Drs. Salvador Sánchez Alonso y Miguel Ángel Sicilia por el apoyo y la orientación recibida incluso antes de formar parte de la Universidad de Alcalá (UAH). Sin su dedicación, aportes y recomendaciones no habría sido posible llegar satisfactoriamente al término de esta tesis. Vuestra capacidad de trabajo y disposición a ayudar los hacen únicos, muchísimas gracias.

También quiero agradecer por su ayuda a mi co-director Manuel Prieto y al grupo de investigación *Information Engineering* de UAH por su invitación a integrarme al equipo, para mi será un honor continuar mi trabajo de investigación con ustedes.

Del mismo modo quiero agradecer a la Universidad del Bío-Bío, a la Facultad de Ciencias Empresariales y al Departamento de Ciencias de la Información, que a través de sus directivos siempre me han brindado el apoyo necesario para cumplir con esta meta.

Capítulo 1. Introducción a la investigación

El objetivo de este capítulo es presentar una síntesis de las áreas relacionadas con esta tesis dando énfasis en aquellos aspectos que motivan su desarrollo. Luego se detalla la forma como se encuentra organizado el resto del documento.

1.1. Introducción

El crecimiento exponencial en la cantidad y diversidad de la información disponible en Internet dificultan la tarea de búsqueda y aumentan la insatisfacción del usuario frente a una lista de cientos o miles de resultados, sin siquiera la certeza de que éstos satisfagan sus requerimientos de información. Dicho problema es conocido como *information overload* (Ho & Tang, 2001). En el caso de los recursos para el aprendizaje la situación si bien no es igual es bastante similar (Ouyang & Zhu, 2007). Hoy en día existe una gran cantidad de recursos dispersos en la red o disponibles en repositorios, por lo tanto las tareas de búsqueda y selección pueden complicarse para el usuario, sea este el aprendiz o el propio diseñador instruccional.

Los esfuerzos dedicados al mejoramiento del proceso de enseñanza–aprendizaje en e-learning son numerosos y persiguen distintos propósitos, por ejemplo:

- Facilitar el diseño y composición de los recursos de aprendizaje.
- Asistir o automatizar el etiquetado de los recursos.
- Facilitar la reutilización de los recursos de aprendizaje, dispersos y distribuidos en diversos repositorios.
- Mejorar y hacer más eficientes las formas de búsqueda y selección de los recursos más apropiados, entre otras.

De acuerdo a las líneas de trabajo antes mencionadas, se encuentra:

- Estándares para el diseño de recursos de aprendizaje (IMS, 2003b), estándares o perfiles de aplicación para el etiquetado de recursos (AENOR, 2008; IMS, 2003b; LTSC, 2002) .
- Plataformas de soporte para las comunidades educacionales o los sistemas de búsqueda federada (Hatala, Richards, Eap, & Willms, 2004a; Morales, Gil, & García, 2007; Ternier, Olmedilla, & Duval, 2005).
- Estándares para el intercambio e interoperabilidad entre repositorios (CEN, 2005; IMS, 2003a).
- Aplicaciones y herramientas para potenciar la interoperabilidad entre repositorios (Bastiaan, Lalanne, & Shamseldin, 2003; Dehors & Zucker, 2006; Hatala, Richards, Eap, & Willms, 2004b; Kaczmarek & Landowska, 2006; Motz, Sosa, & Rodriguez, 2006; Segura N., Menéndez, & Prieto, 2008; Simon et al., 2005).
- Repositorios semánticos (Soto, García-Barriocanal, & Sánchez-Alonso, 2007).
- Entornos inteligentes para el aprendizaje SS4L (Dolog, Simon, Nejdl, & Klobucar, 2008).
- Sistema de recomendación y meta búsqueda de recursos (Prieto M., Menéndez D., Segura N., & Vidal C., 2008; Segura N., Menéndez et al., 2008).
- Herramientas para el reciclaje y etiquetado de recursos para el aprendizaje (Menéndez, Vidal, Urzaiz, & Prieto, 2008; Paulsson & Naeve, 2006).
- Modelos para la evaluación de la calidad de los recursos para el aprendizaje (Segura N., Vidal C., & Prieto, 2008; Vidal, Segura N., & Prieto, 2008; Vidal C., Segura N., Campos G., & Sánchez, 2010).

Nuestra investigación está dirigida hacia el proceso de búsqueda de recursos de aprendizaje en repositorios, específicamente a la aplicación de técnicas de expansión de consultas basadas en ontologías de dominio.

1.2. Estructura del documento

El resto de esta memoria de tesis doctoral esta dividida en las siguientes secciones:

- El segundo capítulo detalla los estudios relativos a los objetos de aprendizaje, los estándares de metadatos, los repositorios, los modelos de conocimiento y la recuperación de información, específicamente la expansión de consultas. Adicionalmente, se resume la revisión y análisis de las propuestas previas relacionadas con la expansión de consultas basadas en ontologías.

- El tercer capítulo presenta los aspectos delimitan el contexto de la investigación, los problemas detectados en la recuperación de objetos de aprendizaje en repositorios y las diferencias entre nuestra investigación y las propuestas previas para la expansión de consultas basadas en ontologías. Por último, se exponen la hipótesis de la investigación, los objetivos de la tesis y el método de trabajo.
- En el capítulo cuarto se describe la propuesta de solución que incluye: los criterios para la expansión de consultas basada en ontologías, la forma como serán abordados los problemas detectados y la estrategia de expansión de consultas basada en ontología de dominio aplicada en el contexto de la búsqueda en repositorios de objetos de aprendizaje.
- El quinto capítulo describe el diseño del experimento de evaluación, las consultas de prueba, la forma de evaluación de la relevancia y las métricas para evaluar la efectividad de la propuesta. Luego se detallan los resultados obtenidos en el repositorio de prueba, la evaluación de la relevancia de los expertos, los resultados de las métricas de desempeño obtenidos para los objetos de aprendizaje recuperados con y sin expansión, y finalmente, se discuten los resultados de las métricas de desempeño.
- En el capítulo 6 se exponen las conclusiones y las futuras líneas de trabajo. Así también se detallan las publicaciones y contribuciones derivadas de la investigación.

Capítulo 2. Estado de la cuestión

A lo largo de este capítulo se ofrece una descripción del estado actual de las investigaciones y propuestas desarrollados en las áreas relacionadas con esta investigación, específicamente el concepto y la evolución de los objetos de aprendizaje, especificaciones de metadatos, repositorios así como el proceso de búsqueda de objetos de aprendizaje en repositorios, modelos de conocimiento y la expansión de consultas basadas en ontologías.

2.1. Concepto y estado actual del área de objetos de aprendizaje

El e-learning es una de las áreas hacia la cual se dirige nuestra investigación. Dentro de este ámbito, nuestros objetivos se enfocan hacia el proceso de búsqueda de recursos de aprendizaje en repositorios. El diseñador instruccional recupera estos recursos para componer otros nuevos o utilizarlos en el diseño en un curso. Suponemos que en la medida que los recursos recuperados sean relevantes para la necesidad de información del profesor, la tarea de diseño requerirá menor esfuerzo y en última instancia, se mejoran las oportunidades para que los alumnos alcancen los objetivos pedagógicos propuestos por el profesor al usar dichos recursos.

Si bien el término e-learning tiene gran cantidad de definiciones, en general se refieren al aprendizaje que utiliza herramientas tecnológicas. En los últimos años se evidencia un incremento muy importante de herramientas y recursos para el e-learning. Las plataformas de aprendizaje han evolucionado desde un conjunto de páginas Web desarrolladas con propósitos específicos hasta sistemas de gestión y mejora del aprendizaje.

Los materiales y recursos ofrecidos en los sistemas de e-learning también han evolucionado y lo han hecho en aspectos tales como: la tecnología aplicada, la complejidad y los objetivos pedagógicos para los cuales fueron diseñados. Hoy en día, existe una gran cantidad y variedad de materiales de aprendizaje disponibles en la WWW pero ello no implica necesariamente un mejor o mayor aprendizaje. Gran parte de estos recursos son redundantes, carecen de formalidad o bien poseen una calidad pedagógica y técnica cuestionable o simplemente desconocida. Como es planteado por (Abbas et al.,

2005) los sistemas e-learning crecieron aisladamente, aumentando la cantidad de materiales disponibles pero con una compatibilidad restringida.

Como ha sido analizado por (Wiley, 2003) cuando el diseñador instruccional crea, actualiza o compone materiales de enseñanza, no necesariamente se siguen pautas o estándares, ya sea tanto para el contenido como para sus descriptores (metadatos). Esta situación no solo complica las tareas de búsqueda y acceso, sino que también dificulta la evaluación de la calidad de éstos y su relevancia respecto a la satisfacción de la necesidad de búsqueda.

2.1.1. Objetos de aprendizaje

El concepto de recurso o material en e-learning se ha formalizado en el concepto de objeto de aprendizaje (OA). Aunque existen diversas definiciones para los OA, en este estudio se utilizará la planteada en (McGreal, 2004). Un OA es "cualquier recurso digital reutilizable que tiene encapsulado una lección o ensamblado un grupo de lecciones en unidades, módulos, cursos e incluso programas".

El propósito de los OA es proporcionar un modelo modular basado en estándares que permita la flexibilidad, independencia de plataforma y la reutilización de contenidos de aprendizaje desplegando un alto grado de control a los profesores y estudiantes.

Los OA permiten organizar el contenido digital en unidades independientes de manera que puedan ser utilizados en diferentes contextos (Paulsson & Naeve, 2006). Bajo este enfoque, los OA se componen y descomponen para formar nuevos OA más complejos, tales como unidades de aprendizaje, lecciones o cursos. En general los OA deben ser fáciles de compartir, reutilizar y combinar (Ouyang & Zhu, 2007).

Cuando se reutiliza un OA, se ensambla o simplemente cuando se prepara un curso se han de tomar en consideración diversos aspectos, por ejemplo:

- Respecto al contenido. Aspectos relacionados con la cantidad, complejidad y densidad de información.
- Respecto a la audiencia. Aspectos relacionados con los conocimientos previos, estilos de aprendizaje, y aspectos sociales, geográficos y culturales.
- Respecto a la tecnología. Aspectos relacionados con los requerimientos mínimos hardware, software o de comunicaciones, compatibilidades, entre otros.

Los recursos se crean para un grupo heterogéneo de alumnos cada uno con sus estilos, ritmos, preferencias, metas y pre-requisitos. Lo anterior no solo implica que tanto la plataforma como el sistema e-learning deben ser flexibles en el flujo de aprendizaje y fáciles de personalizar, sino que los OA deben estar descritos de una manera que facilite su selección. En su estudio (Wenying & Deren, 2006) plantean que el contenido de los materiales disponibles en los sistemas de aprendizaje electrónico no facilitan su correcta selección, la mayor parte de los metadatos que los describen se encuentran vacíos o simplemente mal etiquetados (Cechinel, Sánchez-Alonso, Sicilia, & Sartori, 2009; Segura

N., Vidal, Menéndez, Zapata, & Prieto, 2009; M.-Á. Sicilia, García-Barriocanal, Pages, Martinez, & Gutierrez, 2005).

A grandes rasgos el profesor tiene dos opciones a la hora de desarrollar un objeto de aprendizaje, puede construirlo a partir de cero o modificar un recurso existente. En este último caso, el profesor primero debe buscar el recurso que más se ajusta a sus necesidades y luego, adaptar o componerlo con otros recursos.

La búsqueda de recursos para el aprendizaje puede ser realizada directamente en Internet o en los repositorios de OA. Desde otro punto de vista, la búsqueda puede estar dirigida a recuperar el contenido del recurso, los metadatos o el OA completo (contenido y metadatos). En cualquiera de estos frentes existen consideraciones asociadas, por ejemplo la búsqueda en Internet puede ser fácilmente relacionada con la búsqueda de "contenidos digitales" debido a que la mayor parte de los recursos dispersos en la red no poseen metadatos o éstos no están disponibles. Por su parte, la búsqueda en repositorios debería garantizar que los recursos disponibles son pertinentes pedagógicamente, y en la mayoría de los casos se trata de OA con metadatos (aunque no necesariamente bajo un estándar (McGreal, Adelsberger, Kinshuk, Pawlowski, & Sampson, 2008)).

2.1.2. Metadatos y estándares de metadatos

Cuando un profesor prepara un curso, abstrayéndose que utilice o no un método para el diseño de la instrucción, los pasos seguidos por él se pueden resumir en los siguientes:

1º. Definir el contexto del curso.

2º. Buscar, analizar y seleccionar los OA apropiados.

3º. Componer, adaptar los contenidos o utilizar directamente el OA seleccionado.

4º. Organizar los OA de acuerdo al enfoque instruccional, dando forma al curso. Esto implica la repetición de los primeros pasos hasta que el curso esté terminado.

5º. Dejar disponible el curso para que los estudiantes accedan a los recursos preparados y durante el tiempo que este disponible, el profesor debe monitorear el proceso de aprendizaje.

6º. Una vez terminado el curso, el profesor debe realizar las modificaciones con base en la experiencia y retroalimentación obtenida.

En las tres primeras actividades como mínimo, es fundamental que los OA posean una descripción clara, completa y precisa de sus características, sólo de esta forma podrán ser efectiva y eficientemente buscados, catalogados, adaptados, combinados y almacenados (Vargo, Nesbit, Belfer, & Archambault, 2003). Los estándares de metadatos son especificaciones formales usadas para anotar semánticamente recursos educativos de cualquier tipo. Han sido desarrollados para soportar la interoperabilidad

entre máquinas y el descubrimiento y selección de los recursos por las personas (Al-Khalifa & Davis, 2006).

Según el contexto de aplicación estos estándares pueden ser clasificados en aquellos de propósito general y específico, en la Tabla 1 se ejemplifican algunos en cada caso.

Tabla 1: Estándares de metadatos según el contexto de aplicación.

Estándares de metadatos DE PROPÓSITO GENERAL	DE PROPÓSITO ESPECÍFICO
XML auto descriptivo	Bibliotecas (MARC, METS, MODS, FRBR)
Etiquetas <meta> HTML	Archivos (EAD)
DCMI	Educación (IMS, IEEE-LOM, SCORM, GEM, EdNA)
RDF (metamodelo de metadatos)	Información Geográfica (FGDC, ISO19115)
	Audiovisual: VRA, MPEG7, DIG35
	E-Gobierno: GILS, AGLS
	Perfiles de aplicación (LOM –ES, UKLOM)

Los estándares como IEEE-LOM diseñado por *Learning Technologies Standardization Committee (LTSC)* del *Institute of Electrical and Electronics Engineers* (LTSC, 2002) y Dublín Core *metadata elements* (DMCI, 2003) creado por la DCMI (*Dublin Core Metadata Initiative*), están orientados a describir el contenido y formato de los OA a través de un conjunto de metadatos. El estándar IMS-LRM (*Learning Resource Metadata*) (IMS, 2002), de IMS *Global Learning Consortium Inc,* unifica la información requerida para buscar y usar OA (IMS, 2003a). Dicha especificación se basa en IEEE-LOM 1.0. Por otra parte, el estándar SCORM (ADL, 2003) (*Sharable Content Object Reference Model*) de la iniciativa ADL (*Advanced Distributed Learning*), establece la forma como secuenciar los OA para componerlos en un curso o sesión de aprendizaje, es decir la forma como serán estructurados u organizados los objetos de aprendizaje. Este estándar permite encapsular o empaquetar los objetos para que puedan ser incorporados en cualquier sistema e-learning (Jovanovic, Knight, Gasevic, & Richards, 2006). Finalmente, el estándar IMS-LD (IMS, 2003b) pone el énfasis en el diseño de las actividades de aprendizaje.

Además de estándares propiamente dichos, también se utilizan los perfiles de aplicación que corresponden a un subconjunto de un estándar. En general los perfiles de aplicación pueden:

- Elegir únicamente un subconjunto de los metadatos contemplados por un estándar.
- Refinar la estructura de algunos metadatos contemplados por un estándar.
- Introducir nuevas categorías o nuevos metadatos en las categorías existentes.
- Precisar y especificar el uso y el rol de los metadatos existentes, por ejemplo a través de nuevos vocabularios.

A continuación en la Tabla 2 se listan algunos de los perfiles de aplicación más conocidos.

Tabla 2: Perfiles de aplicación y esquemas de metadatos educacionales basados en IEEE-, IMS y DCMES.

IEEE-LOM (Learning Object Metadata)	IMS – LRM.	DCMES (*Dublin Core Metadata Element Set*)
CanCore[1] (Canadá).	SingCore (Singapur).	DC-Ed (*Dublin Core Metadata for Educational Materials*).
KEM (*Korean Educational Metadata* 2.0).		GEM (*Gateway of Educational Materials*).
LOM-ES (AENOR, 2008), España.		EdNA (*Education Network of Australia*).
CELTS[2] -3.1 (*Chinese E-Learning Technology Standard*) (Xiang, Shen, Guo, & Shi, 2003).		VES (*Virtual European School*).
UKLOM [3]Core. Incluye 23 elementos, en su propio marco de trabajo en metadatos denominado UK *Learning Object Metadata Framework*, del Reino Unido.		EUN (*European Schoolnet*).
		TLF (*The Le@rning Federation SOCCI).*

En general el uso de estándares de metadatos facilita la búsqueda, análisis, recuperación, composición y adaptación de OA. No obstante aún existen barreras para alcanzar mayores niveles de reutilización e interoperabilidad tales como las limitaciones identificadas por (Farance, 2003):

- La característica de opcionalidad de los elementos así como la ausencia, ambigüedad o excesiva flexibilidad de vocabularios, puede provocar problemas no sólo en el llenado de campos sino que también en el procesamiento automático de los OA.
- No todos los estándares definen la correspondencia entre sus elementos con los de otros estándares. Una excepción a esta problemática es propuesta en IMS-LRM ya que éste incorpora una guía de correspondencia con DublinCore y IEEE-LOM.
- En IEEE-LOM existen algunos elementos de datos imprecisos, en donde no se tiene claridad del significado que representa el elemento de datos y cómo éste se relaciona con el contenido. En este estudio se destacan los elementos de datos con más problemas (16 ítems). También se critican los valores de dominio para los elementos, puesto que si bien están definidos se manifiesta explícitamente que pueden ser incorporados otros nuevos.
- Por su parte, según es planteado por (Tello Cáceres, 2007), existe cierta falta de madurez del estándar IEEE-LOM ya que contiene atributos de difícil aplicación. Entre las razones analizadas esta la existencia de campos irrelevantes, campos cuya

[1] http://www.cancore.ca/en/

[2] http://media.cs.tsinghua.edu.cn/~pervasive/projects/e-learning/celtsc.html

[3] http://metadata.cetis.ac.uk/profiles/uklomcore

información es de difícil identificación o bien campos cuya importancia a la hora de describir un recurso educativo no es relevante.

A pesar de las críticas, en la actualidad IEEE-LOM y sus perfiles de aplicación constituyen el cuerpo de metadatos para materiales educativos con mayor madurez y reconocimiento en la comunidad de e-learning internacional (McGreal et al., 2008; Tello Cáceres, 2007; Xiang et al., 2003).

En la Figura 1 se representa el análisis desarrollado por (Xiang et al., 2003) respecto a la relación entre los distintos estándares y sus elementos. El diámetro refleja aproximadamente la cantidad de elementos de datos incluidos en cada estándar.

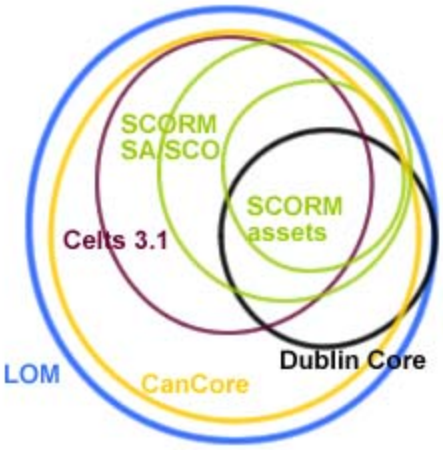


Figura 1: Intersección de los elementos de metadatos. Esquema reproducido desde (Xiang et al., 2003).

El estándar con mayor cantidad de elementos es IEEE-LOM, los siguientes CELTS y CanCore son perfiles de aplicación basados en IEEE-LOM. Por su parte SCORM, en cuanto a los metadatos de los recursos, también se basa en IEEE-LOM. Dublín Core es el estándar de menor alcance aunque esto se debe a que éste corresponde a un modelo general aplicable para describir cualquier recurso.

2.2. Repositorios de objetos de aprendizaje

Como fue mencionado anteriormente, dentro del e-learning nuestros objetivos se enfocan hacia el proceso de búsqueda de recursos de aprendizaje en repositorios. A continuación se describen las principales características y funcionalidades de los repositorios de objetos de aprendizaje (ROA o en inglés *Learning Object Repositories*-LOR). En particular es nuestro interés analizar las opciones y facilidades que estos repositorios proveen al diseñador instruccional para acceder a recursos de aprendizaje según sus necesidades de información y las restricciones de la audiencia hacia donde se destinan dichos recursos.

El estándar IMS -*Digital Repositories Interoperability* (IMS, 2003a), define los repositorios digitales como "una colección de recursos accesibles en la red". Como se representa en la Figura 2, estos depósitos pueden mantener el contenido (a), sus metadatos (b) o ambos (c).

Tras analizar los repositorios disponibles a la fecha, (McGreal et al., 2008) plantea que su estado puede ser descrito como "funcional y creciente" aunque aún no se ha alcanzado la interoperabilidad perfecta de los OA alojados en los distintos repositorios.

Hoy en día existe gran cantidad de repositorios de objetos de aprendizaje[4], este hecho sin duda potencia el usar, compartir, indexar y recuperar recursos, aunque la gran cantidad y diversidad de ellos puede hacer mas compleja la tarea de seleccionar los recursos mas apropiados, es decir que satisfagan los requerimientos del profesor o del aprendiz (Ouyang & Zhu, 2007).

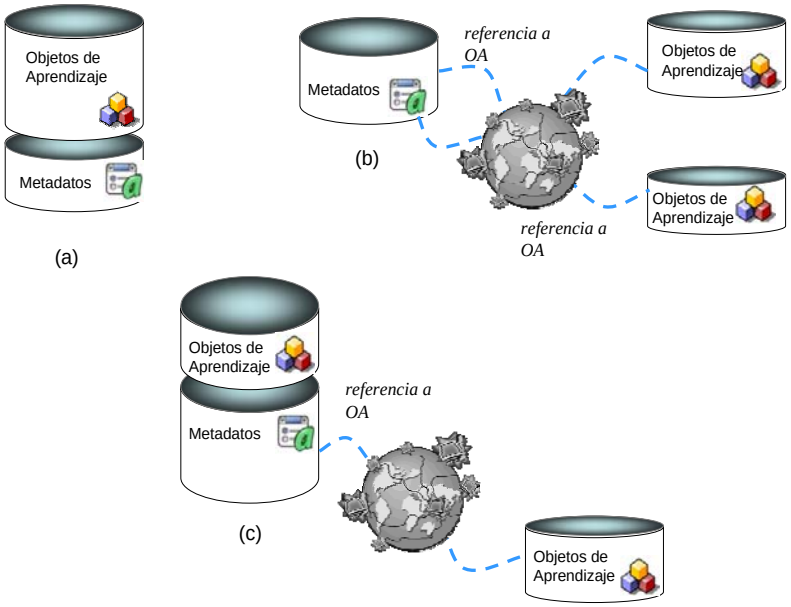


Figura 2: Repositorios según su contenido.

2.2.1. Funcionalidades de los repositorios

IMS-DRI propone una arquitectura funcional y un modelo de referencia para la interoperabilidad entre repositorios (IMS, 2003a). Se definen ocho funciones básicas y cinco combinaciones de ellas como actividades principales:

- buscar/exponer (*search/expose*).
- recolectar/exponer (*gather/expose*).

[4] http://oerwiki.iiep-unesco.org/index.php?title=OER_useful_resources/Repositories, consultado en febrero 2009.

- enviar/almacenar (*submit/store*).
- solicitar/entregar (*request/deliver*).
- alertar/exponer (*alert/expose*).

En este estándar no se establece, define y tampoco se propone la tecnología para implementar dichas funciones, excepto para la función buscar/exponer donde se recomienda usar XQuery, SOAP o Z39.50. Para la función reunir se recomienda el protocolo OAI-PMH (*Open Archives Initiative – Protocol for Metadata Harvesting*) como protocolo de recolección (Kaczmarek & Landowska, 2006). Más detalle al respecto se incluye en la sección 2.2.4.2.

Las funciones definidas en DRI se ubican en dos niveles: de repositorios y de utilización de recursos. A nivel de repositorios se encuentran las funciones de almacenar, exponer, entregar y a nivel de utilización de recursos se encuentran las funciones buscar, enviar, recolectar, alertar y solicitar recursos (Hatala et al., 2004b).

La interoperabilidad entre distintos repositorios, ya sea para el intercambio de OA o sus metadatos, puede ser compleja cuando cada uno de ellos posee representaciones y métodos de acceso heterogéneos. De ahí entonces la importancia de estandarizar los formatos y protocolos de comunicación e intercambio de información. El modelo de referencia de IMS-DRI propone un componente intermedio opcional que puede cumplir una de las tres funciones que simplifican este problema (IMS, 2003a):

- Una función de traductor, puede traducir un formato de búsqueda a otro y es comprendido por múltiples repositorios.
- Una función de agregación que coge los metadatos de múltiples repositorios y los deja disponibles para la búsqueda.
- Una función federada que distribuye una consulta de búsqueda a múltiples repositorios y administra las respuestas.

Algunas de las funciones de los sistemas de búsqueda eficiente definidas en IMS- DRI son: la traducción, la conversión de datos, la distribución de la consulta y los recursos lingüísticos tales como vocabularios, bases de datos léxicas, tesauros, entre otros (Hilera, Bengochea, de Madariaga, de Mesa, & Martínez, 2005) .

Desde el punto de vista de los usuarios administradores o de los contribuidores de recursos, en (Higgs, Meredith, & Hand, 2003) se plantean 9 funciones incorporando aquellas relacionadas a la administración y gestión. Estas funciones son: buscar, controlar la calidad, solicitar OA, mantener versiones, recuperar metadatos, enviar OA, almacenar, recopilar metadatos y publicar metadatos en otro repositorio.

2.2.2. Caracterización de los repositorios

A continuación, se describen las funciones y algunas de las características más relevantes de los repositorios de OA basados en estudios especializados tales como (McGreal et al.,

2008; Neven & Duval, 2002; Ochoa & Duval, 2009; Segura N. et al., 2009; Segura N., Vidal C., Campos G., Menéndez, & Prieto, 2011; M.-Á. Sicilia et al., 2005).

Según la forma como se administran los recursos existen tres tipos de repositorios, aquellos que alojan OA completos (contenido y descriptores), señalados como tipo 1: *contenidos*. Los Repositorios que sólo alojan los metadatos (con enlaces a los contenidos), señalados como tipo 2: *referencias*. Por último, los repositorios que son una mezcla de los dos tipos anteriores, señalados como tipo 3: *mixto*.

Respecto a la clasificación de repositorios el estudio de McGreal et al. (2008), realizado sobre un total de 57 repositorios, concluye que la mayoría de ellos son de *contenido* y una menor cantidad son repositorios de *referencias* o *mixtos*. Esta distribución es comprensible debido al crecimiento y evolución de los objetos de aprendizajes así como de las facilidades de interoperabilidad entre repositorios. No obstante, según (Ochoa & Duval, 2009) el aumento de los mecanismos y estándares de interoperabilidad ha provocado un rápido crecimiento en la cantidad de repositorios de *referencias*.

Según el dominio o área de conocimiento hacia las cuales se dirigen los OA almacenados, se distinguen repositorios dirigidos a áreas específicas o bien repositorios genéricos. Los repositorios en su mayoría son genéricos, independiente del tipo de repositorio al que corresponda (tipo 1, 2 o 3). A pesar de lo anterior, según el estudio realizado por McGreal et al. (2008) los repositorios del tipo 1 tienden principalmente hacia las áreas de Matemáticas, Ciencia y Física. Los repositorios del tipo 2 tienden hacia las áreas de Ingeniería, Lenguaje y Ciencia. Finalmente los repositorios del tipo 3 se inclinan hacia las áreas de las ciencias de la salud, los casos de aprendizaje basado en problemas, el trabajo social y la hidrografía.

Según el nivel educativo al cual se orientan los OA alojados en los repositorios, por ejemplo K-12, primaria, secundaria, universitaria, entre otros. Al respecto, más de la mitad de los repositorios contienen recursos para el nivel universitario, alrededor de un quinto para el nivel de K-12 y también existe un porcentaje importante de los repositorios que alojan OA de todos los niveles.

La granularidad de los OA contenidos en los repositorios también hace distinción entre aquellos que alojan recursos que van desde textos, imágenes y videos hasta otros formatos de recursos más complejos como *Applets* de *java*, animaciones *Flash* o herramientas. Los recursos alojados en los repositorios pueden ser componentes, módulos, lecciones o cursos completos. La mayoría de los repositorios incluyen recursos de variados niveles de granularidad.

Al relacionar el nivel de granularidad asociado a cada uno de los niveles educativos, es posible distinguir que solo para los niveles de primaria y secundaria se observa una especialización al nivel de lecciones, aunque la cantidad de repositorios en estos niveles educativos no es importante (McGreal et al., 2008) .

Finalmente, la cantidad de recursos disponibles en los repositorios varía desde millones como en NDSL *library* hasta unos pocos como en ESOT o SOFIA (Ochoa & Duval, 2009). No obstante esto existe una relación entre la cantidad de recursos y nivel de granularidad, es decir los repositorios más grandes contienen componentes, de menor

grano y los más pequeños acostumbran almacenar cursos completos o aplicaciones especializadas.

A pesar que el estudio de McGreal et al. (2008) se describe que la mayoría de los repositorios implementa algún tipo de control de calidad, también existen otros estudios como en (Cechinel et al., 2009; Ochoa & Duval, 2006; Segura N. et al., 2009; M.-Á. Sicilia et al., 2005) en los que se concluye que gran parte de los metadatos alojados en los repositorios son deficientes, incompletos, inexistentes o simplemente inconsistentes.

Por otro lado, así como fue demostrado en el estudio de McGreal et al. (2008) la mayoría de los repositorios, con un 58%, no utiliza un estándar de metadatos, solo mantienen un conjunto básico de descriptores de cada OA, y en total no más de un 25% utiliza algún estándar formal, refiérase a la Figura 3.

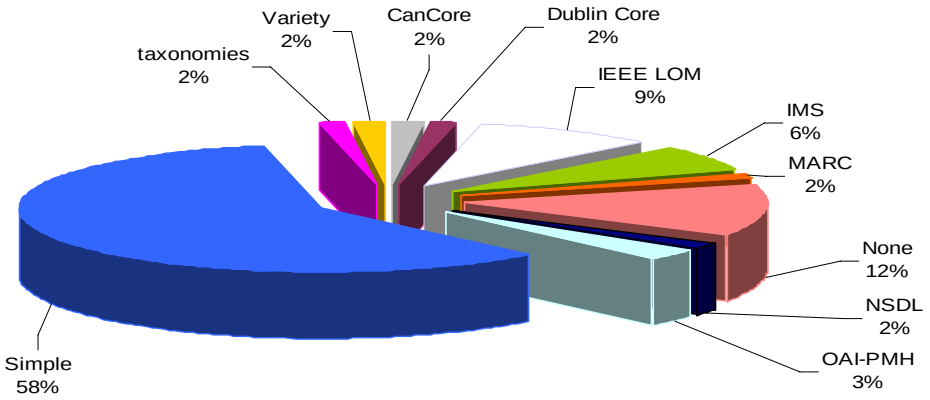


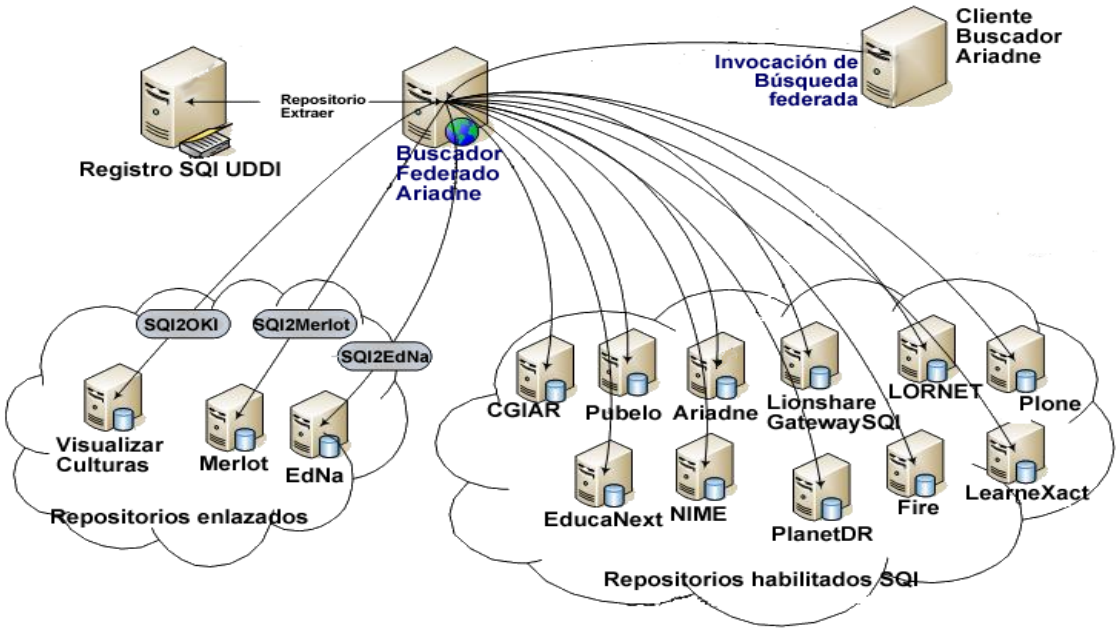
Figura 3: Distribución del uso de estándares de metadatos en los repositorios de OA. Tabla de datos obtenidos de (McGreal et al., 2008).

2.2.3. Interoperabilidad entre repositorios de objetos de aprendizaje

La interoperabilidad entre repositorios puede ser analizada en dos grandes escenarios:

- Búsqueda Federada. La búsqueda federada es implementada a través de un buscador que accede a múltiples bases de datos o repositorios con una misma consulta. Se da cuando una comunidad o grupo de instituciones acuerdan compartir las funciones de búsqueda en sus repositorios, los OA siguen siendo administrados por cada entidad, no obstante pueden ser utilizados por cualquier miembro del grupo. En esta alternativa se requiere un acuerdo en el uso de estándares, esquemas, protocolos y formatos de resultados. Por ejemplo en la Figura 4, se representa la federación de Ariadne.
- Recolección de metadatos. Se basa en proveer un conjunto de servicios que permiten el intercambio de metadatos. Estos se consultan independientemente y

sólo cuando sea necesario se solicitan los recursos (contenidos). Los metadatos recopilados desde los repositorios se almacenan en un repositorio central. Siendo este último quien da la respuesta a las búsquedas. Este sistema permite mayor autonomía, no obstante pueden existir lapsos de tiempo de des-actualización. Es preciso mencionar que no todos los repositorios aceptan "distribuir" sus metadatos, más bien prefieren mantener pleno control sobre sus recursos. En este escenario se utiliza alguna técnica para recopilar la información de los OA desde organizaciones



externas.

Figura 4: Sistema de repositorios federados, ejemplo de la búsqueda federada en Ariadne.

2.2.4. Mecanismos de acceso a los objetos de aprendizaje

En general los repositorios deben mantener los OA y proveer una interfaz de búsqueda que facilite su consulta y recuperación, tanto a los usuarios como a otras aplicaciones. Muchos ROA exponen servicios para el acceso a los OA que almacenan, por ejemplo a través de servicios Web del tipo REST (ver sección 2.2.4.1) u otros servicios de intercambio. La estandarización en el intercambio de OA o metadatos permite la interoperabilidad entre aplicaciones como agentes de búsqueda, sistemas de administración de aprendizaje (*Learning Management System*, LMS), o simplemente otros repositorios de OA.

Entre los estándares utilizados en el escenario de búsqueda federada para soportar la interoperabilidad e intercambio de información entre repositorios podemos mencionar:

- *Search/Retrieve Via Web Service* – SRW y *Search/Retrieve Via Url* - SRU
- *Simple query Interface* - SQI

- *EduSource Commnication layer* - ECL
- *Open Knowledge Innitiative* - OKI

En cuanto a la recolección de metadatos, el protocolo más difundido es *Open Archive Initiative Protocol for Metadata Harvesting* - OAI.PMH.

2.2.4.1 Servicio Web de búsqueda en repositorios

La transferencia de estado representacional (*Representational State Transfer*) o REST es una técnica de arquitectura de software para sistemas hipermedia distribuidos como la *World Wide Web* (Khare & Taylor, 2004).

Si bien el término REST originalmente se refería a un conjunto de principios de arquitectura, en la actualidad se usa en el sentido más amplio para describir cualquier interfaz Web simple que utiliza XML y HTTP, sin las abstracciones adicionales de los protocolos basados en patrones de intercambio de mensajes como el protocolo de servicios Web SOAP. Los sistemas que siguen los principios REST se llaman con frecuencia RESTful.

Un ejemplo es el servicio Web tipo RESTful provisto por el repositorio MERLOT. El uso del servicio Restful MERLOT requiere una clave que se obtiene previa solicitud y registro (MERLOT, 2010). Este servicio permite especificar algunos parámetros opcionales de control y de consulta:

- Parámetros opcionales de control. Por ejemplo el número del registro inicial, cantidad de resultados a recuperar, la opción para recuperar la cantidad total de OA que satisfacen la consulta en el repositorio y especificar el orden de los resultados. En este último caso los valores válidos son: por orden general, por el título, por el autor, por la fecha de creación o por el puntaje promedio de revisión.
- Parámetros opcionales de consulta. Se incluye alguna de las siguientes 3 opciones de búsqueda; cualquiera de las palabras clave, todas las palabras clave o la frase exacta con las palabras clave de la consulta. La consulta puede ser restringida a los campos: título, descripción, autor o derechos de propiedad.

El servicio Restful retorna un archivo XML con los resultados de la consulta. Este archivo se encuentra dividido en 2 partes: un resumen y la lista de resultados. El resumen incluye el total de resultados obtenidos de la consulta en el repositorio, la cantidad de resultados recuperados y el número del último registro recuperado. La lista de los resultados incluye los metadatos de los OA bajo el estándar IEEE-LOM.

2.2.4.2 Open knowledge initiative

La *Open Knowledge Initiative* (OKI project, 2010) desarrolla y promueve un conjunto de especificaciones para facilitar la interoperabilidad e integración entre los componentes de software de distintos entornos, a través de la definición de estándares para *Service Oriented Architecture* (SOA) (Ternier et al., 2006).

OKI construye una arquitectura abierta y extensible donde se especifica la forma cómo los componentes software de un entorno educacional se comunican entre sí y con otros sistemas. El OKI proporciona una API llamada *Open Service Interface Definition* (OSID) por medio de la cual provee definiciones comunes de datos y servicios.

OSID abarca una amplia gama de servicios desde los más genéricos como la autenticación hasta los servicios específicos tales como gestión de cursos y de calificaciones.

2.2.4.3 Búsqueda /recuperación vía servicio Web y vía URL

Search/Retrieve vía URL (SRU) es un protocolo estándar de búsqueda para consultas en Internet, en este caso la petición/consulta va codificada en la misma URL. Las consultas son representadas usando *Contextual Query Language* (CQL), una sintaxis estándar para representar consultas.

Search/Retrieve Web service (SRU vía HTTP SOAP, inicialmente llamado SRW) es un protocolo estándar de búsqueda basado en servicios Web. SRW provee una interfaz SOAP para las consultas, ésta permite el intercambio de información en entornos descentralizados y se encuentra basado en XML. Así como SRU, las consultas son representadas usando CQL. Este protocolo tiene como objetivo integrar el acceso a distintos recursos de una red y promover la interoperabilidad entre bases de datos distribuidas mediante el uso de entornos de trabajo comunes. En este transporte, la solicitud está codificado en XML y envuelto en algunos elementos adicionales específicos de SOAP. La respuesta es la misma que XML SRU por *get* o *post*, pero envuelto en otros elementos específicos de SOAP. Tanto los Estándares para SRW, SRU, y CQL han sido promulgados por *Library of Congress* de los Estados Unidos.

2.2.4.4 Capa de comunicación EduSource

EduSource Communication Layer (ECL) es desarrollado y soportado por LORE *Research Group en la Universidad de Simon Fraser*. ECL es una implementación de la especificación IMS-DRI (ver 2.2) que utiliza SOAP como capa de comunicación (ver Figura 5).

Para facilitar la implementación de ECL se ha implementado un conector, que es una capa intermedia que expone los servicios eduSource en la forma de manejadores *(Handlers)* ocultando la complejidad de la codificación de los mensajes XML y permitiendo la comunicación con otros servicios eduSource. El conector provee una API para conectar un repositorio a la red.

Dado que EduSource pone énfasis en conectarse con otras iniciativas de redes, provee un segundo mediador, *gateway* ECL. Este *gateway* esta modelado después del adaptador que funciona a nivel de red (Hatala et al., 2004b).

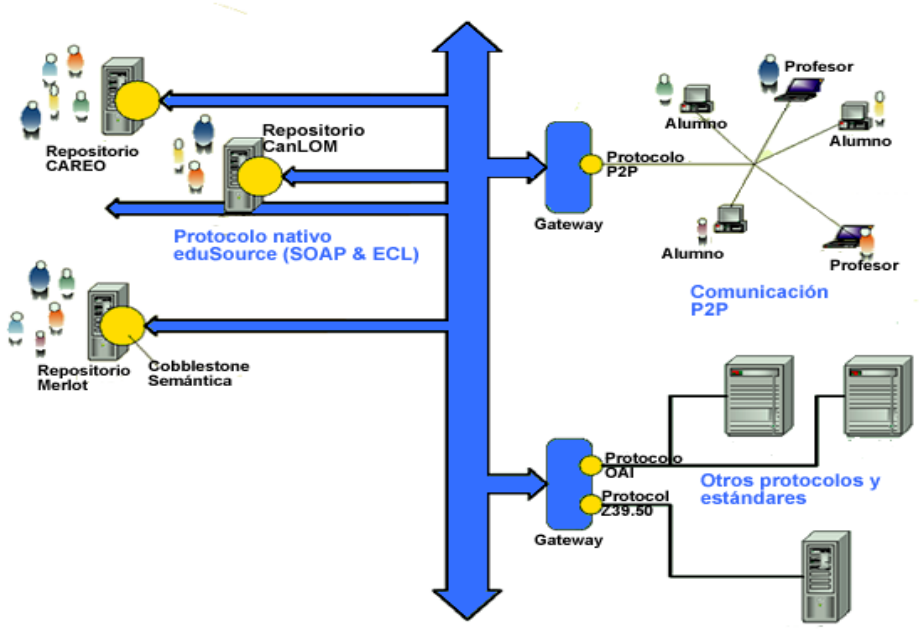


Figura 5: Comunidades soportadas en la infraestructura EduSource. Fuente (Hatala et al., 2004b) pp. 21.

Un lado del *gateway* implementa el conector y por el otro se provee un *framework* definido como una cadena de manejadores que desempeñan la conversión entre el protocolo ECL y los otros protocolos. El *gateway-framework* define la correspondencia entre los protocolos en 4 niveles: el protocolo de comunicación (HTTP, SOAP, XML-RPC, P2P, etc.), el lenguaje de comunicación (ECL, OAI, POOL, etc.), los metadatos (IMS, CanCore, DublinCore), y las ontologías (vocabularios de metadatos).

2.2.4.5 Protocolo para la recolección automática de metadatos

La *Open archive initiative protocol for metadata harvesting* (OAI-PMH) fue creada con el propósito de desarrollar y promover estándares de interoperabilidad para facilitar la difusión eficiente de contenidos en Internet. Esta iniciativa es coordinada por Herbert Van de Sompel y Carl Lagoze, Universidad de Cornell y es financiada por *National Science Foundation*, USA.

Dentro de este protocolo se definen 2 roles participantes (ver Figura 6): el repositorio que soporta OAI.PMH para compartir sus metadatos y los recolectores que usan la recolección de metadatos OAI.PMH para construir repositorios agregados (Lagoze & Van de Sompel, 2002).

OAI-PMH utiliza transacciones HTTP para emitir preguntas y obtener respuestas entre un servidor o repositorio y un cliente o servicio recolector de metadatos. El recolector puede pedir al repositorio que le envíe metadatos según determinados criterios como por ejemplo la fecha de creación de los datos. En respuesta el repositorio devuelve un conjunto de registros en formato XML, incluyendo identificadores de los objetos descritos en cada registro, por ejemplo la URL (Van de Sompel & Lagoze, 2002).

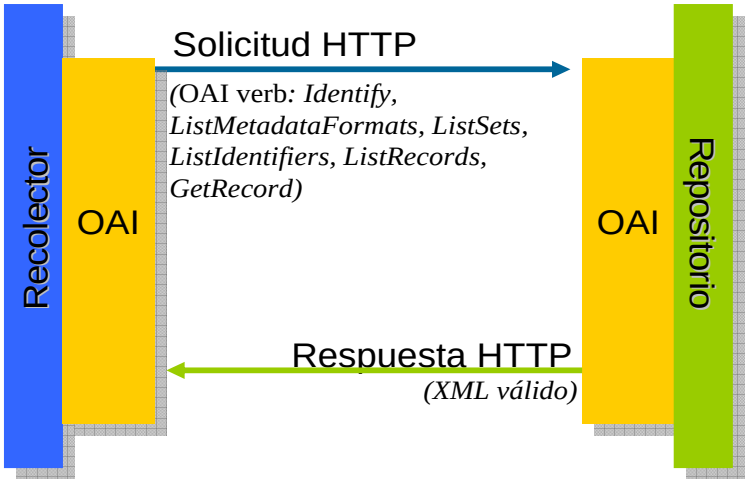


Figura 6: Intercambio de metadatos en El estándar OAI-PMH.

Las peticiones se generan utilizando los métodos *get* o *post* del protocolo HTTP y constan de una lista de opciones con la forma de pares del `tipo:clave=valor`. Existen seis peticiones que un cliente puede realizar a un servidor:

- *GetRecord*. Utilizado para recuperar un registro concreto. Necesita dos argumentos: identificador del registro pedido y la especificación del formato bibliográfico en que se debe devolver éste.
- *Identify*. Utilizado para recuperar información sobre el servidor: nombre, versión del protocolo que utiliza, dirección del administrador, etc.
- *ListIdentifiers*. Recupera los encabezamientos de los registros, en lugar de los registros completos. Permite argumentos como el rango de fechas de los datos que se desea recuperar.
- *ListRecords*. Igual que el anterior pero recupera los registros completos.
- *ListSets*. Recupera un conjunto de registros. Estos conjuntos son creados opcionalmente por el servidor para facilitar una recuperación selectiva de los registros, es decir corresponde a una clasificación de los contenidos según diferentes entradas. Un cliente puede pedir que se recuperen solo los registros pertenecientes a una determinada clase. Los conjuntos pueden ser simples listas o estructuras jerárquicas.

- *ListMetadataFormats*. Devuelve la lista de formatos bibliográficos que utiliza el servidor.

2.2.4.6 Simple query interface

Simple Query Interface (SQI) se define como un conjunto de métodos referidos a una capa de interoperabilidad universal para redes de educación (CEN, 2005; Ternier et al., 2008). En la especificación de SQI no se fija el lenguaje o el formato de los resultados (Simon et al., 2005), por lo tanto la interacción entre un cliente y un proveedor toma lugar cuando se acuerdan entre ellos estos aspectos, refiérase a la Figura 7.

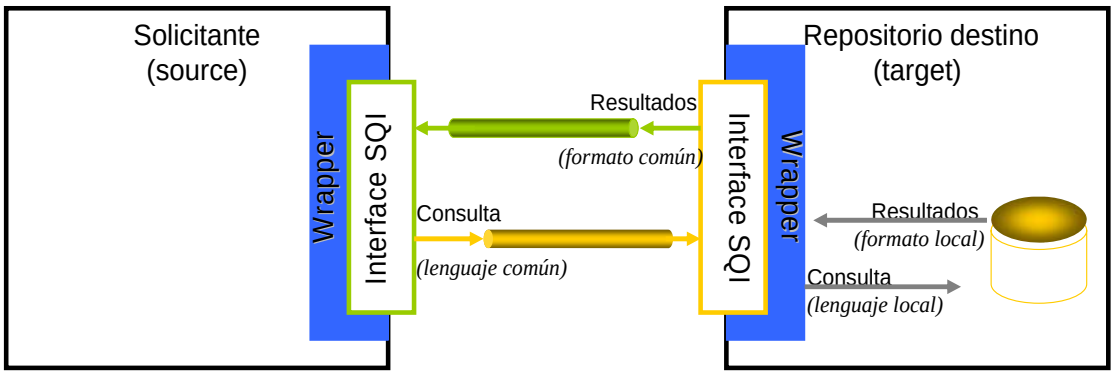


Figura 7: Intercambio de datos en SQI.

Existe una API (*Application Program Interface*) implementada para consultar los OA en repositorios a través de SQI. Considerando que el objetivo es la colaboración entre repositorios muy heterogéneos, se plantean 2 escenarios:

- Síncrono. En este escenario el **destino** devuelve los resultados de la consulta **al solicitante (fuente)**. La recuperación de resultados se inicia desde esta última al enviar la consulta al **destino.**
- Asincrónica. En este escenario la recuperación de resultados es iniciado por el **destino**, una vez que se encuentra una cantidad significativa de resultados. Para soportar esta comunicación, el **solicitante (fuente)** debe implementar un detector de resultados que le permita identificar de forma única una consulta enviada a un **destino** (incluso si la misma consulta se envía a varios destinos).

Servicio Web SQI

Una forma de implementar SQI es a través de servicios Web. Sólo un subconjunto de los métodos propuestos en la API mencionados anteriormente es implementado en esta especificación. El proyecto *Sourceforge*[5] provee un WSDL *binding* básico para construir un

[5] http://sqi-wsdl.sourceforge.net/

servicio Web SQI. Además se incluyen las clases Java *wrapper* para crear e invocar de forma remota el servicio Web vía SOAP usando Apache AXIS. Se implementan 2 tipos de servicios:

La Figura 8 y Figura 9 representan las 2 implementaciones posibles de SQI en Java y SOAP, respectivamente.

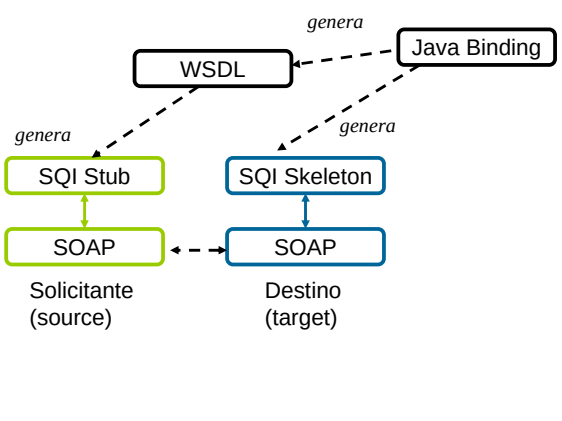


Figura 8: Implementación SQI en Java.

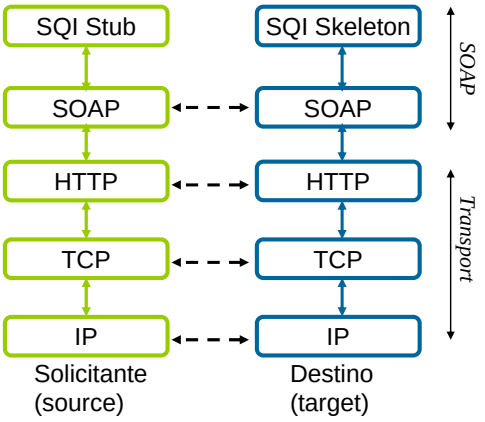


Figura 9: Implementación SQI en SOAP.

Herramienta SQITester

La herramienta SQITester permite acceder a algunos de los repositorios de OA más importantes y que a su vez soportan este estándar (ver Figura 10). La herramienta permite hacer consultas según las especificaciones establecidas por cada repositorio (por ejemplo el lenguaje de consulta, el identificador de la sesión, el formato de resultados, entre otros) y recuperar los metadatos de los OA que coinciden con la consulta en un formato XML compatible con el esquema estándar IEEE-LOM (LTSC, 2002).

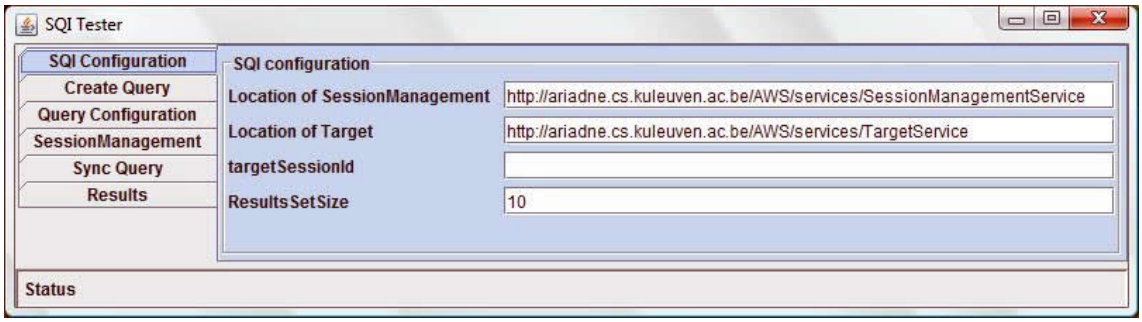


Figura 10. Interfaz SQI tester.

Para comenzar, la **fuente** debe crear una conexión con el destino, ya sea a través de una identificación o con una sesión anónima. Es posible configurar la consulta a través de los siguientes métodos: *setQueryLanguage, setResultsSetSize, setMaxQueryResults, setMaxDuration, setResultsFormat.*

La consulta puede acceder a través de dos métodos de consulta, *asynchronousQuery* o *synchronousQuery*. El lenguaje de consulta más utilizado por los repositorios es VSQL, éste es un lenguaje básico, un mayor detalle al respecto es entregado en 2.2.5. El esquema de resultado más comúnmente soportado por los repositorios es IEEE-LOM.

2.2.5. Lenguajes de consultas en repositorios

Existen dos formas de abordar la búsqueda: exacta y aproximada. La búsqueda exacta se realiza por medio de la búsqueda sobre metadatos, usando motores de búsqueda basados en XML, y la búsqueda aproximada basada en palabras clave se realiza como una búsqueda libre tanto en el contenido textual como en los metadatos, implementada usando motores de recuperación de información tradicionales por ejemplo como Lucene (Ternier et al., 2008). Lucene es una API para la recuperación de información, inicialmente implementada en Java (Sarnowski & Kessel, 2009).

Los lenguajes de consulta tienen el propósito de dar soporte al intercambio de consultas entre repositorios heterogéneos.

- Very Simple Query Language

  Very Simple Query Language (VSQL) es un lenguaje básico que no considera la búsqueda en los campos de metadatos, es decir los metadatos son tratados como un texto cualquiera. Hoy en día, según el registro de repositorios de SQI[6] la mayoría de los servicios de búsqueda soportan este lenguaje para el intercambio de consultas.

  Dada su simpleza este lenguaje es bastante utilizado en los repositorios, aunque también es muy restrictivo debido a que las consultas se implementan sólo como búsquedas aproximadas conjuntivas, es decir que el texto es buscado tanto en el contenido textual como en cualquiera de los metadatos. Además VSQL sólo soporta implícitamente el conector lógico AND, es decir no permite el uso de otros tipos de conectores básicos tales como OR ó NOT.

  Por ejemplo, la consulta "test learning" se interpreta como (test AND learning) y queda expresadas como se observa a continuación.

```
<simpleQuery>\n\n
<term>test</term>
<term>learning</term>
</simpleQuery>
```

- Prolearn query language

  El lenguaje *Prolearn query language* (PLQL) se basa en otros lenguajes como Xquery, CQL (Morgan, 2004) y VSQL (Ternier et al., 2008). Este lenguaje se encuentra especificado en 5 capas o niveles, cada uno de los cuales posee un poder de expresión mayor. PLQL es un lenguaje pensado para la búsqueda en metadatos

[6] Registro SQI, http://ariadne.cs.kuleuven.be/SqiInterop/free/SQIImplementationsRegistry.jsp, consultado en 2 de febrero de 2009

aunque no asume el uso de algún estándar específico. Es importante destacar que a la fecha [7] sólo existen implementaciones en el nivel 0 y 1.

A continuación se describen los aspectos generales de los tres primeros niveles (cuya especificación se encuentra disponible a la fecha):

- Nivel 0: Representa la búsqueda aproximada conjuntiva. Los términos son buscados tanto en el contenido texto como en los metadatos, por lo tanto este nivel es equivalente con el lenguaje VSQL. El ranking de los resultados queda en manos del repositorio aunque por lo general se basa en la relevancia de las palabras claves. En este nivel el lenguaje es independiente a un estándar de metadatos.

- Nivel 1: Da respuesta a la necesidad de extender las consultas con la opción de especificar criterios de selección por ejemplo relacionados con rangos de edad, nivel educativo, lenguaje, entre otros. En este nivel se mezcla tanto la búsqueda aproximada como la exacta. Las cláusulas exactas se ejecutan primero, luego sobre el conjunto de resultados obtenidos se aplica la búsqueda aproximada y se calcula el ranking.

  En los casos en que la búsqueda exacta no retorne resultados es posible especificar un comportamiento alternativo, por ejemplo trasformar la búsqueda exacta en una búsqueda aproximada. Este nivel sólo soporta la conjunción y la igualdad como predicados de comparación. Hasta la fecha sólo el repositorio MACE (Stefaner et al., 2007) ha implementado la búsqueda tanto aproximada como exacta para el acceso externo a los OA.

  A continuación se presenta un ejemplo de consulta exacta donde se especifica el contenido de los campos de metadato requeridos, en este caso se requieren OA en cuya descripción contengan el texto "*multiple Choice test*", cuyo *nivel de agregación* sea 1 ó 2, que permita un tipo "*active*" con un *nivel* "medium" de *interactividad*, cuya *dificultad* de operación sea "*easy*", con *densidad semántica* "*medium*" y específicamente un recurso de *tipo* "*exam*".

```
(lom.general.description.string = "multiple choice
tests") and
(lom.general.aggregationLevel = "1" or
lom.general.aggregationLevel = "2" ) and
(lom.educational.interactivityType= "active") and
(lom.educational.interactivityLevel = "medium") and
(lom.educational.difficulty = "easy") and
(lom.educational.learningResourceType = "exam") and
(lom.educational.semanticDensity = "medium")
```

- Nivel 2: Toma en cuenta la naturaleza jerárquica de los metadatos de objetos de aprendizaje. En las cláusulas exactas es posible usar disyunciones y todos los predicados convencionales de comparación.

[7] Revisado al 2 de febrero de 2009, http://ariadne.cs.kuleuven.be/SqiInterop/free/SQIImplementationsRegistry.jsp

2.2.6. Formato de los resultados recuperados del repositorio

La mayoría de los servicios de consulta de OA o recolección de metadatos entregan los resultados, por defecto, según el estándar de metadatos IEEE-LOM.

En la especificación de PLQL se definen 4 niveles de resultados, cada uno con mayor detalle. A continuación se describen los cuatro niveles definidos para el formato de los resultados:

Nivel 0. En este nivel los resultados sólo incluyen la cantidad de resultados recuperados.

Nivel 1. En este nivel los resultados incluyen la URI y, si el proveedor utiliza un método de ranking los resultados, deben retornarse ordenados de mayor a menor relevancia.

Nivel 2. En este nivel se agrega información de los metadatos de los resultados. No existe una restricción respecto de cuáles y cuantos campos se retornarán. Por lo menos se asume que en los resultados se incluye el título, autor y lenguaje.

Nivel 3. En este nivel se agrega a los resultados el puntaje de relevancia de cada resultado y el método de ranking utilizado.

2.2.7. Análisis de las funcionalidades de búsqueda en repositorios

Para conocer detalladamente las opciones y funcionalidades de búsqueda que ofrecen los repositorios al usuario se han analizado 13 de los repositorios más importantes según la cantidad de recursos que contienen o hacen referencia (McGreal et al., 2008). Específicamente se incluyen los repositorios NSDL, CITIDEL, INTUTE UK, GEM, MERLOT, Edna, DLESE, BioDiTRL, SchoolNet Canadá, SMETE, ARIADNE – GLOBE, Learn Alberta, y CAREO (Neven & Duval, 2002). La evaluación[8] realizada se concentra en las opciones y funcionalidades de búsqueda que están disponibles para el usuario y el procesamiento lingüístico realizado, ya sea la eliminación de *stopword*[9] y la extracción de la raíz de los términos de la consulta (refiérase a la Tabla 3). Para evaluar la eliminación de *stopword* se compararon los resultados al probar consultas con y sin *stopword*. Un proceso similar se realizó en el caso de la extracción de la raíz[10], al comparar resultados de palabras con raíces comunes, por ejemplo *debug* y *debugging*, *studies* y *study*, entre otras.

[8] Evaluación realizada en marzo 2010.

[9] Entiéndase *stopword* como aquellas palabras que son generales, comunes y que no agregan información respecto al contenido de un documento, más detalle se encuentra descrito en la sección 2.4 página 37.

[10] Entiéndase que la forma canónica de la raíz de las palabras representa las variaciones de los términos derivados de ella, más detalle se encuentra descrito en la sección 2.4 página 37.

Tabla 3: Síntesis de la evaluación realizada a las opciones de búsquedas provistas en los repositorios.

Repositorio		**Opciones de búsqueda**	**Elimina *Stopword***	**Extrae raíz**
http://nsdl.org/search/	NSDL	Búsqueda por palabras clave, provee opciones para restringir por grado, formato, nivel y la colección de recursos. La búsqueda especializada permite refinar los filtros mencionados. Permite la expresión de consultas utilizando conectores lógicos.	SI	SI
http://www.citidel.org/	CITIDEL	Incluye búsqueda avanzada a partir de la cual es posible seleccionar el campo de búsqueda entre abstract, palabras clave, título, tema, serie, patrocinador y lenguaje. Es posible componer una expresión lógica a través de conectores AND, OR o NOT. Permite la navegación por áreas de conocimiento y por comunidad, así como refinar la búsqueda mediante un vocabulario controlado.	SI	SI
http://www.intute.ac.uk/	INTUTE UK	Incluye la búsqueda avanzada a partir de la cual es posible seleccionar opciones de filtro como tema y tipo de recurso. Es posible elegir entre los campos de búsqueda título, descripción y palabras clave. Incluye opciones para adaptar la salida, por ejemplo ordenar alfabéticamente o por relevancia, así también permite listar todos los campos o solo el título. Cuando no se recuperan resultados existe la opción de ampliar la búsqueda utilizando un cliente de búsqueda de Google, el cual en vez de buscar en la Web completa lo hace solo en el catálogo. Permite la expresión de consultas utilizando conectores lógicos.	SI	NO*
http://www.thegateway.org	GEM	La búsqueda puede ser limitada a los campos título, palabras claves o descripción. Ofrece la búsqueda a través de la navegación entre categorías para los campos tema, tipo, nivel, mediador, palabras clave, beneficiario y precio. Se incluye la navegación por los conceptos buscados más frecuentemente. Permite la expresión de consultas utilizando conectores lógicos.	SI	SI
http://www.merlot.org	MERLOT	Ofrece la búsqueda simple, avanzada y distribuida. En su opción de búsqueda avanzada permite buscar por varios campos a la vez, por ejemplo palabras clave, título, URL, descripción, comunidad, categoría de tema, lenguaje, tipo material, formato técnico, audiencia, sistema administrador de aprendizaje LMS. Además permite encontrar recursos por autor, por comunidad contribuyente o por rango de fechas.	SI	NO
http://www.edna.edu.au/edna/go	Edna	Provee la opción de búsqueda estándar, avanzada y distribuida. En la segunda permite la búsqueda a través de los campos tales como; palabras clave, título, audiencia, creador, descripción, tema, editor y categoría Edna. Es posible filtrar por tipo de recurso y sector. Otra opción es la selección de los tesauros que pueden ser utilizados. Permite la expresión de consultas utilizando conectores lógicos.	SI	SI
http://www.dlese.org/dds/histogram.do?group=subject	DLESE	Además de la búsqueda simple, ofrece la búsqueda a través de la navegación entre categorías grado, tema, tipo de recurso y estándares. Es posible ordenar y filtrar los resultados. Permite la expresión de consultas utilizando conectores lógicos.	SI	SI
http://bio-ditrl.sunsite.ualberta.ca/	BioDiTRL	Ofrece búsqueda simple, y según los resultados la búsqueda se realiza en campos título y descripción.	SI	SI
http://www.schoolnet.com	SchoolNet Canada	Ofrece opción de búsqueda simple.	NO	SI
http://www.smete.org/smete/	SMETE	Búsqueda por título, grado, tipo, fecha de publicación, autor y palabras clave. Permite modificar el orden de los resultados por título, fecha o relevancia. Permite la expresión de consultas utilizando conectores lógicos.	SI	SI

Repositorio		Opciones de búsqueda	Elimina *Stopword*	Extrae raíz
http://www.globe-info.org/en/search/node	ARIADNE - Globe	Búsqueda simple y avanzada, permite la selección de los repositorios que serán consultados entre Edna, Lornet, MINE, Merlot. La búsqueda avanzada permite especificar el tipo de concatenación de los términos de búsqueda como todas las palabras, alguna o la frase exacta.	SI	NO
http://www.learnalberta.ca/	Learn Alberta	Búsqueda simple, permite filtrar los resultados según algunos campos como grado, tema, audiencia, idioma, formato o tipo de recurso. Permite la expresión de consultas utilizando conectores lógicos.	SI	SI
http://www.ucalgary.ca/commons/careo/search.htm	CAREO	Implementa una interfaz de búsqueda distinta, al parecer la búsqueda se realiza en el título o descripción. Permite establecer si se desea coincidir mayúsculas o si busca la palabra completa.	NO	NO

* Si la búsqueda se amplía con el cliente Google, si se realiza extracción de la raíz.

La Figura 11 resume los resultados de la evaluación realizada.

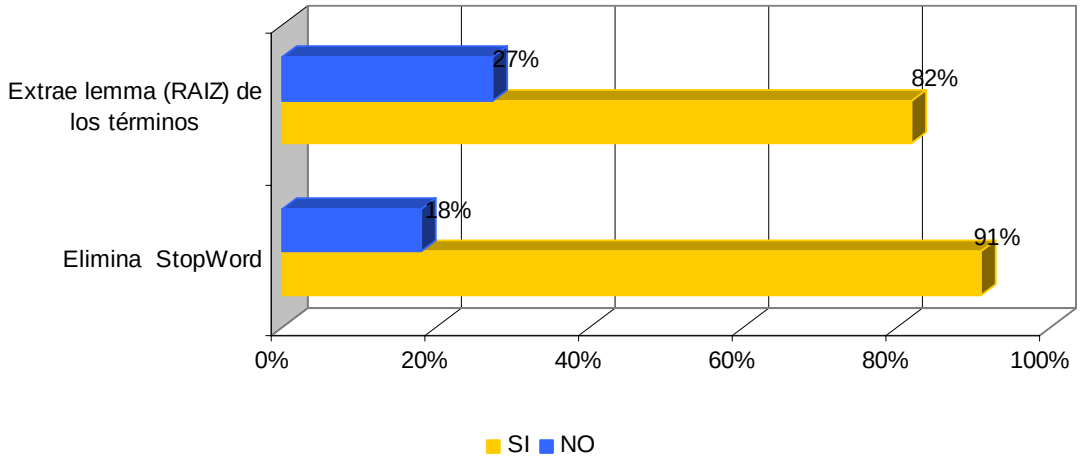


Figura 11: Resultados de la evaluación de la forma de búsqueda utilizada en los repositorios.

Tras la prueba y observación realizada a las funciones de búsqueda se puede mencionar:

- La mayoría utiliza la búsqueda por palabras clave en el título, área, o descripción del OA, y en menor cantidad se incorporan otros descriptores.
- Se utilizan campos de filtrado, los más frecuentes son el nivel educativo, el tipo de recurso o el área de conocimiento. Si bien en cada campo se utiliza un vocabulario este no siempre es estándar.
- Otra opción es la navegación en jerarquías de categorías o subcategorías de tópicos.
- Una opción, aunque menos frecuente, permite configurar la cantidad de campos de salida y el orden de los resultados.
- Es frecuente que las palabras *stopword* sean eliminadas de los términos de la consulta.

- Es menos frecuente la extracción de la raíz de los términos de las consultas, es decir que se simplifique o amplíe la búsqueda según la agregación de prefijos o sufijos a los términos. Por ejemplo, si la consulta es *debug* en la búsqueda se deberían incluir los términos *debugging y debuger*.

Respecto a los resultados de las búsquedas y considerando el acceso de un usuario sin registro o privilegio, se obtuvo:

- La opción de descarga del contenido original es escasa, sólo en casos de recursos texto es posible descargar en formato *pdf*. No se encuentran disponibles las opciones para la descarga de los metadatos o del OA empaquetado.
- La lista de resultados de la búsqueda solo incluye descriptores generales, tales como título, autor, formato, nivel educativo o fecha de ingreso/actualización. En muy pocos casos es posible acceder a otros campos como la descripción, palabras clave, estado del OA, nivel de interactividad, dificultad, entre otros.
- En el caso de Multibúsqueda, búsqueda distribuida o federada los resultados incluyen la referencia a la fuente desde donde se obtuvo el OA y la posición en el ranking de los resultados.

2.3. Representación del conocimiento

Como ya hemos señalado, en nuestra investigación se integran distintas áreas entre ellas la representación de conocimiento. Específicamente es de nuestro interés el uso de los modelos de conocimiento en el proceso de búsqueda de recursos de aprendizaje en repositorios. Por lo tanto, a continuación se describen los aspectos generales en esta área y en particular se analizan las características y diferencias de las ontologías respecto a otros modelos de conocimiento. Dicha información sirve de base para analizar el uso de estos modelos en el proceso de expansión de consultas.

La representación de conocimiento se preocupa de la forma como representar los hechos del mundo real con un cierto formalismo. El grado de formalidad de esta especificación puede ir desde muy informal a rigurosamente formal. En la medida que las descripciones posean mayor formalidad es posible que un sistema informático las utilice (Marr, 1982; Rich & Knight, 1991).

Para lograr la representación del conocimiento hace falta un lenguaje de representación, la capacidad de inferencias y el conocimiento del dominio. Según Winston una representación es "…un conjunto de convenciones sintácticas y semánticas que hacen posible el describir cosas" (Winston, 1992). De ahí que una representación debe poseer:

- Capacidad lógica para expresar el conocimiento deseado.
- Poder heurístico para inferir y resolver problemas a partir del nuevo conocimiento.
- Conveniencia de la notación, de manera que se facilite el entendimiento y acceso al conocimiento.

Como lo indican (M.-A. Sicilia, García-Barriocanal, Sánchez-Alonso, & Soto, 2005), las ontologías son una buena herramienta para la representación formal del conocimiento orientada al análisis semántico por parte de máquinas. De aquí su coherencia con la visión de la Web Semántica expresada por (Berners-Lee, 1998) como "disponer datos en la Web definidos y enlazados de forma que puedan ser utilizados por las máquinas no solamente para visualizarlos sino también para automatizar tareas, integrar y reutilizar datos entre aplicaciones".

2.3.1. Ontologías

En inteligencia artificial, las ontologías aportan el lenguaje de comunicación necesario en entornos distribuidos, éstas se tratan como descripciones para que un sistema informático las utilice.

Las ontologías involucran dos partes: una sintaxis y una semántica. La primera considera los símbolos y el conjunto de reglas para combinarlos, y la segunda se refiere al significado de las expresiones construidas. A continuación se mencionan algunas de las definiciones dadas en la literatura al concepto Ontología.

- Según (Chandrasekaran, Josephson, & Benjamins, 1999), las ontologías son "los modelos de representación de conocimiento que permiten un alto nivel de conceptualización formal del conocimiento, así como también permiten que éste conocimiento sea fácilmente compartido".
- Según (Neches et al., 1991), "Una ontología define los términos y las relaciones básicas para la comprensión de un área, así como las reglas para combinar los términos para definir las extensiones del vocabulario".
- Según (Borst, 1997), "Una ontología es una especificación formal de una conceptualización compartida".
- Según (Weigand, 1997), "Una ontología es una base de datos que describe conceptos generales o específicos sobre un dominio, algunas de sus propiedades y cómo los conceptos se relacionan unos con otros".
- Según (Uschold, Benjamins, Gómez-Pérez, Guarino, & Robert, 1999), "Una ontología necesariamente incluirá un vocabulario de términos y una especificación de su significado (definiciones e interrelaciones entre conceptos) que impone estructura al dominio y restringe las posibles interpretaciones".
- (Lamarca Lapuente, 2006) sintetiza las definiciones dadas por Gruber, Neches, Borst y Weingand, y plantea que "Una ontología es un sistema de representación del conocimiento que resulta de seleccionar un dominio o ámbito del conocimiento, y aplicar sobre él un método con el fin de obtener una representación formal de los conceptos que contiene y de las relaciones que existen entre dichos conceptos".

Una de las definiciones más aceptadas es la propuesta por (Gruber, 1993), "*Una ontología es una especificación formal y explícita de una conceptualización compartida*". A partir de esta definición se extraen algunos aspectos clave como:

- **especificación formal** indica que el conocimiento modelado se representa a través de un lenguaje formal e interpretable fácilmente, es decir que el conocimiento queda expresado formalmente a través de un conjunto de símbolos que poseen un significado.
- **especificación explícita** hace mención a la necesidad de describir todos y cada uno de los conceptos que conforman la ontología.
- **conceptualización compartida** se refiere a que la ontología representa una visión consensuada de la realidad en relación a un contexto de uso.

Independiente del área en el cual se utilice, las ontologías permiten:

- Representar y compartir el conocimiento utilizando un vocabulario común.
- Usar un formato de intercambio de conocimiento.
- Establecer un protocolo específico de comunicación.
- Reutilizar el conocimiento.

Las ontologías representan formalmente especificaciones de conceptos que ofrecen un conocimiento compartido en un dominio definido sobre un lenguaje semántico. La estructura de una ontología se compone de:

- Conceptos (clases y subclases).
- Propiedades que describen las clases y subclases. Las propiedades describen relaciones entre tipos de objetos individuales. Un objeto pertenece a una clase y se relaciona con otros objetos.
- Restricciones o características de una propiedad.
- Taxonomía de los conceptos o relaciones formales entre conceptos.
- Instancias, que representan objetos determinados dentro de un concepto y axiomas, que permiten inferir mediante reglas el conocimiento que no está explícitamente indicado en la taxonomía de conceptos.

Las ontologías especifican rigurosamente un esquema conceptual en un dominio, con el objetivo de facilitar la **comunicación**, la **interacción**, el **intercambio** y el **compartir información** entre diferentes sistemas computacionales.

2.3.1.1 Clasificaciones

Las ontologías pueden ser clasificadas según distintos criterios entre los cuales se pueden mencionar su aplicabilidad, punto de vista, nivel de estructuración, ámbito de conocimiento, tipo de usuario, grado de abstracción, entre muchos otros. En la Tabla 4

se describen algunas de las clasificaciones propuestas por (Devedziz, 2006; Guarino, 1997; Hernández & Sáiz N, 2007; Lamarca Lapuente, 2006; Steve et al., 1998; Van Heijst, Schreiber, & Wielinga, 1997). Como será posible notar existen algunas clasificaciones cuyas categorías son claramente dicotómicas, aunque también hay clasificaciones con categorías que no son excluyentes entre si.

Tabla 4: Tipos de ontologías según distintos criterios de clasificación.

Criterio de clasificación	Categorías
Según el ámbito de conocimiento al que se apliquen:	Ontologías generales. Son las ontologías de nivel más alto ya que describen conceptos generales (el espacio, el tiempo, la materia, etc.).
	Ontologías de dominio. Describen el vocabulario de un dominio concreto del conocimiento.
	Ontologías específicas. Son ontologías especializadas que describen los conceptos para un campo limitado del conocimiento o una aplicación concreta.
Según el tipo de usuario al que vayan destinadas:	Ontologías lingüísticas. Estas se vinculan a aspectos lingüísticos, es decir a aspectos gramáticos, semánticos y sintácticos destinados a su utilización por los seres humanos.
	Ontologías no lingüísticas. Estas están destinadas a ser utilizadas por robots y agentes inteligentes.
	Ontologías mixtas. Son aquellas donde se combinan las características de las anteriores.
Según el asunto que conceptualizan:	Ontologías de un dominio. Son ontologías en las que se representa el conocimiento especializado pertinente de un dominio o subdominio, como la medicina, las aplicaciones militares, tráfico etc.
	Ontologías genéricas. Ontologías en las que se representan conceptos generales y fundacionales del conocimiento como las estructuras parte y todo, la cuantificación, los procesos o los tipos de objetos, independientes de un dominio en particular.
	Ontologías representacionales. Ontologías en las que se especifican las conceptualizaciones que subyacen a los formalismos de representación del conocimiento, por lo que también se denominan meta-ontologías (*meta-level* o *top-level ontologies*).
	Ontologías de tareas. Ontologías creadas para una actividad o tarea específica.
	Ontologías de aplicación. Ontologías creadas para una aplicación específica.
Según su uso en el e-learning (Devedziz, 2006; Hernández & Sáiz N, 2007).	Ontología de dominio: describe los conceptos esenciales, relaciones y teorías de los diferentes dominios de interés.
	Ontología de tareas: los conceptos y relaciones que se incluyen en este tipo de ontología pertenecen a los tipos de problemas, estructuras, partes, actividades y pasos a seguir en el proceso de solución de problemas.
	Ontología para la estrategia de la enseñanza: provee instructores y actores con la facilidad de modelar experiencias en la enseñanza, especificando el conocimiento y los principios de las diferentes acciones pedagógicas y comportamientos.
	Ontología de modelo de aprendizaje: se utiliza para construir modelos y es esencial para los sistemas que representan escenarios de aprendizaje adaptativo.
	Ontología de interfaz: especifica el comportamiento adaptativo y las técnicas en el nivel de interfaz de usuario.
	Ontología de comunicación: se utiliza en el intercambio de mensajes entre las diferentes plataformas, repositorios y servicios educativos. Define la semántica en que se basarán los mensajes, por ejemplo, el vocabulario de términos que se utilizarán en la comunicación.
	Ontología de servicios educacionales: Este tipo se encuentra estrechamente relacionada con la ontología de comunicación. Está basada en OWL-S y proporciona los medios para crear descripciones, procesables por computadores y sistemas.

Criterio de clasificación	Categorías
Según el nivel de abstracción y el razonamiento lógico que permitan:	Ontologías descriptivas. Aquellas que incluyen descripciones, taxonomías de conceptos, relaciones entre los conceptos y propiedades, pero no permiten inferencias lógicas.
	Ontologías lógicas. Aquellas que permiten inferencias lógicas mediante la utilización de una serie de componentes como por ejemplo la inclusión de axiomas, etc.
Según la cantidad y tipo de la conceptualización:	Ontologías terminológicas. Son aquellas que especifican los términos que son usados para representar conocimiento en un área. Suelen ser usadas para unificar vocabulario en un dominio determinado.
	Ontologías de información. Son aquellas que especifican la estructura de almacenamiento de bases de datos. Ofrecen un marco para el almacenamiento estandarizado de información.
	Ontologías de modelado del conocimiento. Son aquellas que especifican conceptualizaciones del conocimiento. Contienen una rica estructura interna y suelen estar ajustadas al uso particular del conocimiento que describen.

2.3.1.2 Lenguajes ontológicos

Simple HTML Ontology Extensions (SHOE), es uno de los primeros lenguajes ontológicos para la Web. A través de SHOE es posible incorporar en un documento HTML una serie de etiquetas con información semántica. Su problema estaba en la forma cómo se etiquetaban los contenidos, y el escaso poder expresivo que se podía alcanzar (Samper Zapater, 2006).

Al poco tiempo la W3C publica el *Resource Description Framework* (RDF), que pasaría a ser la base de la mayoría de los lenguajes ontológicos de la actualidad. RDF fue definido como una herramienta de modelado de metadatos basado en la sintaxis XML. Es decir que XML actúa como un lenguaje para modelar datos y RDF como un lenguaje para especificar metadatos.

RDF es una definición para formar tripletas con un **asunto** y un **objeto** vinculado por una **propiedad** o **predicado** que los relaciona o los vincula de alguna manera (ver Figura 12).

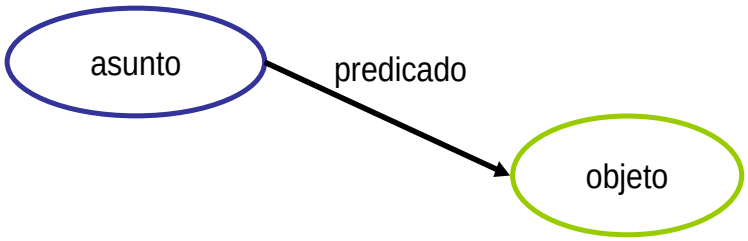


Figura 12: Representación de una Tripleta RDF.

Cada tripleta representa una declaración de una relación entre los artefactos denotados por los nodos que enlaza. Sus tres componentes son:

- Un sujeto que es una referencia URI de RDF o un nodo blanco (es decir, un nodo que no esta identificado por una URI y tampoco es un literal).

- Un predicado que es una referencia URI de RDF (también llamado una propiedad) el cual denota una relación.
- Un objeto que es una referencia URI de RDF, un literal o un nodo blanco.

Por ejemplo, la expresión *Santiago es la capital de Chile*, puede ser representada:

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-
ns#" xmlns:dc="http://purl.org/dc/elements/1.1/">
 <rdf:Description rdf:about="Santiago">
 <dc:escapital>Chile</dc:escapital>
 </rdf:Description>
</rdf:RDF>
```

A partir del lenguaje RDF surgen otras tecnologías que amplían su poder expresivo con el fin de poder representar de la forma más potente posible el conocimiento de cada dominio, a continuación en la Tabla 5 se presenta una síntesis con los lenguajes ontológicos más utilizados.

Tabla 5: Características relevantes de los principales lenguajes ontológicos. Síntesis desde (McGuinness & van Harmelen, 2004; Motik, Patel-Schneider, & Parsia, 2009; Samper Zapater, 2006).

	Nombre	Definición general	Características más importantes
SHOE	*Simple HTML Ontology Extensions.*	Primer lenguaje de etiquetado para diseñar ontologías en la Web. Este lenguaje nació antes de que se ideara la Web Semántica.	Permite definir clases y reglas de inferencia.
OIL	*Ontology Inference Layer.*	Derivado en parte de SHOE.	Se basa tanto en la lógica descriptiva (declaración de axiomas) y en los sistemas basados en frames (taxonomías de clases y atributos). Posee varias capas de sub-lenguajes, entre ellas destaca la capa base que es RDFS, a la que cada una de las capas subsiguientes añade alguna funcionalidad y mayor complejidad.
OWL	*Ontology Web Language*	Lenguaje de etiquetado semántico para publicar y compartir ontologías en la Web. Es una extensión del lenguaje RDF y emplea sus tripletas, aunque es un lenguaje con más poder expresivo que éste.	Puede usarse para representar ontologías de forma explícita y formal (lógica descriptiva), es decir, permite definir el significado de términos en vocabularios y las relaciones entre éstos términos. Se trata de un lenguaje diseñado para el procesamiento automático. OWLv2 añade nuevas funcionalidades con respecto a OWLv1. Algunas de las nuevas características ofrecen nueva expresividad, incluyendo: • *key* • las cadenas de propiedad • tipos de datos enriquecidos, rangos de datos • restricciones de cardinalidad calificadas • propiedades asimétricas, reflexivas, y disjuntos, • mejoras en las capacidades de anotación

	Nombre	Definición general	Características más importantes
KIF	*Knowledge Interchange Format*	Es un lenguaje para representar ontologías basadas en la lógica de primer orden. Fue creado con el objetivo de actuar como interlingua entre diferentes formalismos y lenguajes de representación. KIF dispone de su propia sintaxis y algunos añadidos semánticos sobre la lógica de primer orden.	Basado en la lógica de predicados posee extensiones para definir términos, meta-conocimiento, conjuntos, razonamientos no monotonitos, etc. Pretende ser un lenguaje capaz de representar la mayoría de los conceptos y distinciones actuales de los lenguajes más recientes de representación del conocimiento. Se trata de un lenguaje diseñado para intercambiar conocimiento entre sistemas de computación distintos, diferentes lenguas, etc.
DAML y OIL	*DARPA's Agent Markup Language y Ontology Inference Layer*	El lenguaje DAML se desarrolló como una extensión del lenguaje XML y de RDF, para extender el nivel de expresividad de RDFS.	Se aleja del modelo basado en clases (frames) y potencia la lógica descriptiva. Es más potente que RDFS para expresar ontologías. Ofrece un rico conjunto de elementos con los cuales se pueden crear ontologías y marcar la información para que sea legible y comprensible por una máquina. También funciona como formato de intercambio.

En la Tabla 6 se presenta una comparación de los lenguajes ontológicos obtenida del análisis realizado en (Gómez-Pérez, Fernández-López, & Corcho, 2004) y actualizada para este estudio. En ella se han destacado en color azul los lenguajes DAML+OIL y OWL como aquellos lenguajes más potentes y completos en cuanto a las facilidades que ofrecen para expresar el conocimiento.

Tabla 6: Comparativa Lenguajes Ontológicos. Análisis inicial obtenido desde (Gómez-Pérez et al., 2004) y actualizado por el autor.

Ítem de comparación		**XML DTD**	**XML Schema**	**RDF(S)**	**DAML+OIL**	**RDF(S) 2002**	**OWL 2**
Listas limitadas	Se refiere a la posibilidad de indicar el final de listas o colecciones.				X	X	X
Restricciones de Cardinalidad	Se refiere a limitar el número de declaraciones con el mismo tema y predicado	X	X		X		X
Expresiones de Clase	Se refiere a proveer otras expresiones que involucran unionOf, disjointUnionOf, intersectionOf, o complementOf.				X		X
Tipos de datos	Se refiere a si se incorporan otros tipos de datos (además del String).		X		X	X	X
Clases definidas	Se refiere a si permite expresar restricciones o expresiones asociadas a las clases.				X		X
Enumeraciones	Se refiere a si permite especificar un conjunto de valores para un atributo.	X	X		X		X
Equivalencia	Se refiere a si soporta la equivalencia y el razonamiento.				X		X
Extensibilidad	Se refiere a si soporta que las nuevas propiedades puedan ser utilizadas en las clases existentes.			X	X	X	X
Semántica Formal	Se refiere a si permite la expresión de la semántica a través de modelos teóricos y formas axiomáticas.				X	X	X

Ítem de comparación		XML DTD	XML Schema	RDF(S)	DAML+OIL	RDF(S) 2002	OWL 2
Herencia	Soporta la definición de herencia.			X	X	X	X
Inferencia	Soporta inferencia en todos los niveles.				X		X
Restricciones locales	Permite definir restricciones a relaciones clase – propiedad.				X		X
Restricciones Calificadas	Permite definir restricciones calificadas. Por ejemplo "a lo más x", "todos", entre otros.				X		X
Rectificación	Permiten que una declaración sea tema de otra declaración.			X	X	X	X

Según (McGuinness & van Harmelen, 2004), OWL proporciona un lenguaje para definir ontologías estructuradas que pueden ser utilizadas a través de diferentes sistemas, usuarios, bases de datos o cualquier aplicación que necesita compartir información específica. Las ontologías incluyen definiciones de los conceptos básicos en un área de conocimiento y la relación entre ellos.

OWL se presenta en tres versiones (Samper Zapater, 2006; W3C, 2008):

- OWL LITE: Es un subconjunto sintáctico de OWL DL, es un lenguaje más fácil de entender o implementar ya que restringe o prohíbe el uso de ciertos constructores y axiomas. Permite crear jerarquías de clasificación y restricciones sencillas. Posse baja expresividad pero se supone mayor eficiencia y además ofrece una forma sencilla de migrar un Tesauro o Taxonomía al formato OWL.
- OWL DL: Permite un alto nivel de expresividad sin perder por esto la completitud computacional de los sistemas de razonamiento. Posee una semántica bien definida y algoritmos de razonamiento. Se denomina así ya que permite representar la lógica descriptiva (DL), un subconjunto de la lógica de primer orden.
- OWL Full: Posee mayor expresividad que OWL DL pero no garantiza que se puedan realizar razonamientos en tiempos computables. Permite expresiones de lógica de segundo orden.

Cada sub-lenguaje es una extensión del anterior, por lo que cualquier ontología OWL Lite es también OWL DL y OWL Full, sin embargo, lo contrario no es aplicable.

Las versiones de OWL LITE y OWL DL corresponden a lenguajes de lógica descriptiva, por lo que permiten su procesamiento mediante sistemas de razonamiento. La lógica descriptiva (LD) proporciona un lenguaje formal para combinar y construir definiciones de categorías (taxonomías), así como algoritmos eficientes para decidir las relaciones de subconjunto y superconjunto entre categorías. La base de conocimiento definida en LD incluye dos partes, la definición de términos o conceptos **(TBOX)** y la descripción de individuos mediante aserciones (**ABOX**).

Los sistemas de razonamiento son un conjunto de axiomas más reglas de inferencia que ofrecen propiedades como **consistencia** –todo lo que se deduce es correcto,

completitud –todo lo que es correcto puede deducirse, con **decisión** –existe un algoritmo para decidir si se cumple una conclusión, **expresividad** –capacidad de representar un problema, y **tratable** –el algoritmo de decisión tiene una complejidad razonable.

Las principales tareas de inferencias en LD son:

- Comprobar si una categoría es un subconjunto de otra por medio de la comparación de sus definiciones (*Subsunción*).
- Comprobar si un objeto pertenece a una categoría (*Clasificación*).
- Comprobar si el criterio de pertenencia a una categoría puede ser satisfecho lógicamente (*Consistencia*).

Entre los *razonadores* más conocidos se encuentran:

- *Renamed Abox and Concept Expression Reasoner*, RACER (Volker Haarslev & Möller, 2001; V. Haarslev, Möller, & Turhan, 1998; V. Haarslev, Möller, & Turhan, 1999; Racer Systems GmbH & Co. KG, 2009).
- Jena2 (Reynolds, 2010).
- *Multi-version Ontology REasoner*, MORE (WASP Center Server, 2006).
- Pellet (Clark & Parsia, 2010; Sirin, Parsia, Grau, Kalyanpur, & Katz, 2007).
- FACT++ (School of Computer Science- University of Manchester, 2010).
- Hoolet (Department of Computer Science at University of Manchester, 2004).

2.3.2. Tesauros

Los tesauros también son otra forma de representar el conocimiento. Según lo señala la norma ISO 2788 un tesauro es "un vocabulario controlado y dinámico, compuesto por términos que tienen entre ellos relaciones semánticas y genéricas, y que se aplica a un dominio particular del conocimiento" (ISO, 1986).

Los tesauros tienen un conjunto preestablecido de relaciones ya sean de equivalencia, jerárquicas y asociativas. Por lo general, estas relaciones no cambian según el dominio. Algunas de ellas son:

- *broader term* (término genérico). Se utiliza el operador BT y en español se utiliza el operador TA.
- *narrower term* (término específico). Se utiliza el operador NT en español se utiliza el operador TE.
- *used for* (usado por). Se utiliza el operador UF.
- *related term* (término asociado o véase también). Se utiliza el operador RT, en español se utiliza el operador TR.

- *scope note*. Representa una nota de aplicación, se utiliza el operador SN.
- *Microthesaurus*. Representa una relación de inclusión en un microtesauro, se utiliza el operador MT.

Para efectos de ejemplificar, mas adelante se presenta un extracto del tesauro ETB-LRE MEC-CCAA V.1.0. Este tesauro se ha desarrollado en el seno del grupo de trabajo 9 perteneciente al subcomité 36 de tecnologías de la información y la comunicación para el aprendizaje (SC36) de la Asociación Española de Normalización y Certificación (AENOR) al que pertenecen las Administraciones Públicas Estatales y Autonómicas que lo editan. En la Tabla 7 se presenta un ejemplo para el término *educación especial*.

Tabla 7: Parte de la especificación de términos y relaciones en el tesauro ETB-LRE MEC-CCAA.

ES: educación especial	**Tipo de relación**
CT: educación especial	**idioma Catalán**
EN: special education	**idioma Inglés**
EU: hezkuntza berezia	**idioma Eusquera**
GA: educación especial	**idioma Gallego**
MT: 04 sistema de enseñanza –	**microthesaurus**
SN: tipo especial de educación para niños excepcionales, principalmente para deficientes mentales o físicos	**scope note**
UF: pedagogía terapéutica	**used for**
BT1: sistema educativo	**broader term**
NT1: formación para la vida cotidiana	**narrower term**
NT1: educación compensatoria	**narrower term**
RT: excepcional	**related term**
RT: deficiente	**related term**
RT: escuela en hospitales	**related term**
RT: escuela de educación especial	**related term**
RT: profesor de educación especial	**related term**
RT: literatura grabada	**related term**

Por ejemplo Tesauro ETB-LRE_MEC-CCAA_V.1.0 en formato XML queda expresado como:

```
<term>
     <termIdentifier>23</termIdentifier>
     <caption>
           <langstring>adolescencia</langstring>
     </caption>
</term>
<term>
     <termIdentifier>24</termIdentifier>
     <caption>
           <langstring>adolescente</langstring>
     </caption>
</term>
<relationship>
     <sourceTerm>23</sourceTerm>
     <targetTerm>24</targetTerm>
     <relationshipType source="ETB-LRE MEC-
CCAA">TR</relationshipType>
</relationship>
```

2.3.3. Comparación entre ontologías y tesauros

Si bien las ontologías son similares a los tesauros, existen algunas diferencias fundamentales entre ellos. Éstas se hayan en el nivel de abstracción, en las relaciones entre conceptos, en la capacidad para que el conocimiento representado sea comprensible para las máquinas, en la formalidad y en la capacidad de expresividad que pueden proporcionar (Baader, Calvanese, McGuinness, Nardi, & Patel-Schneider, 2007; Ding & Foo, 2002a, 2002b; Gómez-Pérez et al., 2004; Kwasnik, 1999; Jian Qin & Paling, 2001).

En las ontologías los tipos de relación son arbitrarios, diversos y específicos al dominio, ya que pueden extenderse según las necesidades de especificación del área de conocimiento que se desea representar. En cambio en los tesauros los tipos de relaciones están pre-establecidos y no varían según el dominio en el que se utilice.

Las ontologías son más complejas que los tesauros puesto que permiten llegar a un mayor nivel de profundización semántica y proporcionan una descripción lógica y formal que puede ser interpretada tanto por las personas como por las máquinas. Normalmente se considera que un tesauro es un caso particular de ontología, en tanto que el primero posee una capacidad expresiva considerablemente menor.

En las ontologías el grado de formalización es mucho mayor que en los tesauros. Las ontologías pretenden describir el mundo (o cuando menos un dominio) sobre la base de una lógica descriptiva (a través de OWL). La incorporación de axiomas sobre las clases, las relaciones y las propiedades (de simetría, transitividad, equivalencias, etc.) hacen posible razonar formalmente sobre ellas. Lo anterior también implica mayores costos para formalizar el conocimiento en una ontología.

Las ontologías son modelos más legibles ya que se representan en lenguajes transportables, por ejemplo OWL, RDF, entre otros (McGuinness & van Harmelen, 2004). Por lo tanto, también es mayor la capacidad de reutilización y de distribución del conocimiento a través de sistemas heterogéneos.

2.4. Recuperación de información

La última área de conocimiento a tratar dentro de la revisión del estado del arte es la recuperación de información, en particular es de nuestro interés revisar y analizar las técnicas de expansión de consultas basadas en modelos de conocimiento. A partir de dicha información será posible evaluar la aplicabilidad de estas técnicas en la recuperación de objetos de aprendizaje en repositorios especializados.

La recuperación de información (en adelante RI) involucra una serie de procesos que pueden variar en distintos contextos, si bien en general se distinguen la indexación, la consulta, la evaluación y la retroalimentación del usuario (Baeza-Yates & Ribeiro-Neto, 1999).

- Proceso de indexación. Se encarga de crear y mantener actualizada una base de datos con todos los documentos que pueden ser recuperados. Cabe destacar que la forma como se indexe será muy importante a la hora de consultar y acceder a los documentos.
- Proceso de consulta. La consulta se realiza a través de alguna herramienta. Para la búsqueda de información en Web, se distinguen los directorios o índices temáticos, los buscadores, los multi y metabuscadores.
- Proceso de evaluación y de retroalimentación del usuario. Este proceso consiste en determinar la relevancia de los documentos recuperados para luego presentarlos de manera que reflejen su importancia para el usuario. La retroalimentación de relevancia o la información asociada a la interacción entre el usuario y el sistema, puede ser obtenida explícita o implícitamente. La primera requiere que el usuario evalúe la importancia o pertinencia de cada resultado según su necesidad de información. En la segunda, el sistema de recuperación puede aprender o deducir la relevancia de los resultados a partir de la interacción con el usuario, por ejemplo a partir de resultados seleccionados, revisados o descargados, orden de selección, tiempo de revisión, entre otros. Con lo anterior, se intenta reducir la carga cognitiva requerida del usuario durante la retroalimentación de relevancia y reducir el tiempo necesario asociado a esta evaluación.

Dentro del proceso de recuperación de información, la indexación corresponde a la transformación del documento en una representación que facilite su búsqueda y acceso. Las técnicas de indexación deben identificar *buenos descriptores del documento* en cuanto a que reflejen su contenido y *buenos discriminadores,* es decir que faciliten la diferenciación de un documento de los demás documentos de la colección.

Para indexar lo primero que se realiza es transformar el contenido textual del documento a minúsculas, eliminar caracteres especiales y puntuaciones. Luego, se eliminan aquellas palabras que son generales, comunes y que no agregan información respecto al contenido del documento, en inglés estas palabras son denominadas *stopwords*. El conjunto de estas palabras esta compuesto por preposiciones, artículos, adverbios, conjunciones, pronombres posesivos y demostrativos, algunos verbos y sustantivos. Las listas de *stopwords* se definen dependiendo del idioma, aunque también existen algunas listas creadas para un dominio o área específica de conocimiento (Zazo, Figuerola, Alonso Berrocal, & Emilio, 2005). Aún después de eliminar los *stopwords* no todos los términos que restan pueden ser utilizados en la indexación. Una forma de reducir esta cantidad es realizar un proceso de extracción de la raíz de los términos (en inglés definido como *stemming*). El *stemming* o lematización es el proceso por el cual las variaciones morfológicas de los términos, generados por la agregación de prefijos y sufijos, son eliminadas dejando sólo la raíz de los términos. La forma canónica de la raíz representa las variaciones de los términos derivados de ella (Porter & Sparck-Jones, 1997). Por lo tanto después del *stemming* se simplifican las palabras que posean la misma raíz. Los algoritmos de stemming utilizan distintas técnicas, ya sean basadas en reglas o en diccionarios. Uno de los algoritmos de lematización basado en reglas más sencillo es el lematizador S (Hull & Grefenstette, 1996). Este algoritmo se limita a eliminar las

terminaciones plurales de los términos. El algoritmo de lematización más famoso es el algoritmo de (Porter, 1980; Porter & Sparck-Jones, 1997). Dicho algoritmo elimina cerca de 60 terminaciones en cinco etapas. En cada etapa se extrae un tipo concreto de terminación, eliminándolo o transformando la raíz. También se destacan los algoritmos de (Lovins, 1968), (Paice, 1990), y los algoritmos basados en diccionario, como KSTEM (Krovetz & Um, 1993).

Finalmente, después del *stemming*, se determinan los términos que mejor representan al documento en la indexación. Los términos menos frecuentes entre los documentos de la colección son *mejores discriminantes*. La determinación del peso de los términos de un documento a indexar se realiza considerando la importancia de éstos: en la colección, en el documento individual, o en una combinación de ambos.

La frecuencia inversa del documento (*inverse document frequency, idf*) y la frecuencia del término (*term frequency, tf*) son dos de las medidas más utilizadas para determinar el peso de los términos. La frecuencia inversa del documento (Ecuación 1) pesa el término en función de su frecuencia inversa en la colección, por lo tanto a menor frecuencia del término en la colección implica un mayor peso.

$$idf(t) = \ln \frac{N}{n} \qquad \text{Ecuación 1}$$

Donde:
N : cantidad de documentos en la colección.
n: cantidad de documentos que contienen el término t.

Por su parte la medida de frecuencia de término (Ecuación 2) pesa el término en función de su frecuencia en el documento, es decir asigna un mayor peso a los términos con una frecuencia mayor en el documento.

$$tf(t) = \ln \frac{(occs_t)}{\ln(length_d)} \qquad \text{Ecuación 2}$$

Donde:
$occs_t$: cantidad de ocurrencias del término t en el documento d.
$length_d$: cantidad de términos en el documento d.

El proceso de indexación define los términos a través de los cuales el sistema de recuperación de información podrá acceder a los documentos, obviamente que en este proceso se pierde mucha información respecto al contenido del documento. Hoy en día es posible mejorar notablemente esta situación por medio de modelos de conocimiento que representen la semántica (Baeza-Yates & Ribeiro-Neto, 1999).

Así como es mencionado por (French, Powell, Gey, & Perelman, 2001), "una de las tareas más importantes pero frustrantes en la recuperación de información es la formulación de la pregunta". Es muy difícil expresar en unas cuantas palabras una

necesidad de información, más aún considerando la ambigüedad del lenguaje natural en lo relativo a la expresión de la consulta y a la expresión del contenido textual de los recursos.

El proceso de búsqueda se origina con un problema que requiere información para resolverse. En general, los sistemas de RI asumen que las necesidades de información pueden describirse o expresarse en la forma de una petición. La petición es una representación de la necesidad de información del usuario en un lenguaje humano, normalmente lenguaje natural (Mizzaro, 1996). Cuando esta petición se expresa en un lenguaje comprensible a los sistemas de RI se habla de consulta.

Normalmente el usuario expresa su necesidad de búsqueda utilizando un conjunto de palabras que se denomina **cadena de búsqueda** y opcionalmente, podrá combinarlas utilizando algunos operadores de búsqueda. Esta cadena es modificada por el sistema de RI según sus características particulares, por ejemplo en cuanto al lenguaje, sistema de indexación, índices, uso de operadores, entre otras. El resultado de estas transformaciones es una expresión denominada **términos de búsqueda**.

En (Muramatsu & Pratt, 2001) se mencionan algunas de las transformaciones más comunes de la cadena de búsqueda:

- Aplicación de un operador booleano de búsqueda por defecto.
- Eliminación de palabras *stopwords*.
- Expansión de los términos con sufijos (términos que compartan la raíz).
- Influencia del orden de los términos en el resultado de la búsqueda.

Cuando el usuario conoce las características y transformaciones que realiza un sistema RI puede adaptar su consulta de forma conciente para mejorar los resultados. No obstante esta información no es de conocimiento general para los usuarios.

En los sistemas de RI, tradicionalmente, las búsquedas consideran la palabra de una manera aislada del resto del texto. Algunos de los problemas relacionados con este tratamiento son la sinonimia, polisemia, homonímia, hiperonímia, hiponímia, entre otros. En general todos estos problemas se traducen en que un mismo concepto pueda expresarse con palabras diferentes y una misma palabra pueda aparecer en documentos que tratan sobre temas distintos.

Uno de los desafíos en la recuperación de información se refiere a dar respuesta a preguntas tales como: "*¿que quiere el usuario?, ¿para que necesita la información?, ¿que espera obtener?*", es decir conocer el significado de la consulta del usuario. La naturaleza ambigua de las palabras y la gran cantidad de resultados de las búsquedas, son aspectos que pueden ser abordados con un enfoque semántico, es decir, dándole significado y sentido a las palabras dentro de un contexto.

La semántica estudia el significado de los signos lingüísticos y de sus combinaciones. El orden de las palabras, la frase o la concatenación de palabras de la cadena de búsqueda entregan información valiosa sobre el significado, dominio y propósito de la consulta del usuario. El significado de una misma palabra depende del contexto lingüístico que la

envuelve y que determina su significado. Debido a este fenómeno, se pueden distinguir dos tipos de significado: referencial y contextual. Una palabra tiene un significado referencial cuando se "*refiere a*" su relación convencional con la realidad, estos significados se pueden encontrar en un diccionario. En cambio, el significado "*contextual*" es el que adquiere la palabra dentro de un contexto, cuando amplía, restringe y aún transforma su significado referencial (Muramatsu & Pratt, 2001).

En el contexto de las consultas no solo intervienen factores lingüísticos, sino también sociales y culturales. De ahí la complejidad asociada a determinar y restringir en contexto de la consulta del usuario para mejorar la relevancia de los resultados recuperados.

El proceso de recuperación se traduce en establecer la correspondencia entre la consulta del usuario y el índice de documentos. Dependiendo de la forma como se confronta la consulta y los documentos indexados se distinguen distintos modelos de recuperación tales como booleano, espacio-vectorial, probabilístico y lógico (Baeza-Yates & Ribeiro-Neto, 1999).

2.4.1. Métricas de desempeño de los sistemas recuperación de información

Las métricas más comunes para evaluar el desempeño de los sistemas de RI son la precisión y la exhaustividad (*recall*) (Baeza-Yates & Ribeiro-Neto, 1999; Borlund, 2003; Dolog et al., 2008; Jarvelin & Kekalainen, 2000; Voorhees, 2001). La capacidad de un sistema de RI para proveer de documentos relevantes se mide a través de la métrica de exhaustividad (en inglés *recall*). Esta métrica, descrita en la Ecuación 3, evalúa la proporción de material relevante recuperado respecto del total de los documentos que son relevantes en la colección, independientemente de que éstos se recuperen o no.

$$\text{recall} = \frac{\text{CRrel}}{\text{CRrel Colección}} \qquad \text{Ecuación 3}$$

Donde:
`CRrel`: es la cantidad de resultados relevantes recuperados.
`CRrelColeccion`: la cantidad total de resultados relevantes de la colección.

Dado que cualquier sistema de recuperación de información podría conseguir un 100% de exhaustividad simplemente devolviendo todos los elementos de la colección, se utiliza también la métrica de precisión (descrita en la Ecuación 4). La precisión evalúa la proporción de material recuperado relevante respecto del total de los documentos recuperados.

Con el tiempo han surgido variaciones o mejoras a las métricas clásicas de precisión y exhaustividad, no obstante éstas siguen siendo las más utilizadas. Entre otras razones esto se debe a que muchas de las nuevas métricas aún no han sido ampliamente probadas sobre distintas colecciones estándar (Demartini & Mizzaro, 2006).

$$\text{precición} = \frac{\text{CRrel}}{\text{TR}} \qquad \text{Ecuación 4}$$

Donde:
`CRrel`: es la cantidad de resultados relevantes recuperados.
`TR`: la cantidad total de resultados recuperados.

2.4.2. Ranking de relevancia en los sistemas de recuperación de información

Los algoritmos de evaluación que utilizan muchos de los buscadores actuales se basan en la estructura de la Web para determinar su relevancia. Estos algoritmos de evaluación se denominan algoritmos basados en enlace y son tres principalmente; PageRank, HITS (*Hypertext Induced Topic Selection*) y SALSA (*Stochastic Approach for Link Structure Analysis*).

Los algoritmos basados en enlace se apoyan en la estructura de la Web, considerada como un grafo dirigido de páginas y enlaces. Una página con muchos enlaces entrantes se supone que es una página de alta calidad, especialmente si los enlaces vienen de páginas que son a su vez de alta calidad. Por tanto, se puede considerar a la Web como un grafo dirigido `G=(P,E)` donde `P` son los nodos o páginas Web y `E` son los enlaces entre las páginas.

Los algoritmos de este tipo sufren el "efecto de la contribución circular". Este efecto se basa en el hecho de que las páginas se pueden enlazar unas a otras, de forma que se produzca un camino circular entre ellas. Por tanto, cada página estimula la evaluación de las páginas a las que se enlaza, y si existe un camino circular, entonces estimula su propia evaluación indirectamente. Como una forma de evitar este problema, en (Wang, Wang, & Lee, 2004) se propone la aplicación del concepto de "distancia en la Web" de forma que se asignen pesos a los enlaces en función de la importancia de la página enlazada (Lempel & Moran, 2000).

Los algoritmos HITS y SALSA son específicos a un tema y se pueden considerar algoritmos de evaluación locales. Estos dos algoritmos funcionan utilizando una pequeña porción de la Web donde es probable que existan los recursos correspondientes de un tema específico, analizan la estructura de enlaces de ese sub-grafo Web y asignan a sus páginas puntuaciones concentrador (*hub)* y autoridad. Una página es una autoridad en un tema si contiene información valiosa y de alta calidad sobre ese tema. Una página es un concentrador si enlaza a buenas autoridades sobre el tema.

2.4.2.1 Algoritmo de ranking PageRank

El algoritmo PageRank define un camino aleatorio con saltos aleatorios sobre la Web (completa). Los estados del camino aleatorio son las páginas Web, y la puntuación de cada página se define mediante sus valores de distribución estacionarios del camino aleatorio. Es decir que la puntuación PageRank de una página se puede interpretar como

global, evaluando la importancia de cada página independiente del tema (Lempel & Moran, 2005).

La puntuación PageRank de una página `X` (denotada como `PR(X)`) es la probabilidad de visitar `X` a través de un camino aleatorio que implique a toda la Web. Cada paso aleatorio es de uno de los siguientes tipos:

- Elegir una página Web aleatoriamente, y saltar a ella.
- Desde un estado `S` dado, elegir aleatoriamente un enlace saliente de `S` y seguir ese enlace hasta la página destino.

En (Brin & Page, 1998)se describe el cálculo del algoritmo PageRank a través de la siguiente ecuación:

$$PR(X) = (1-d) + d\left(\frac{PR(T_1)}{C(T_1)} + \ldots + \frac{PR(T_n)}{C(T_n)}\right)$$ Ecuación 5

Donde:
`X`: es una página con las páginas $T_1 \ldots T_n$ que apuntan a ella.
`d`: es la probabilidad de que un visitante que navega en una página se aburra de ella y solicite otra. Puede tomar valores comprendidos entre `0` y `1`, normalmente se establece `d` con el valor `0,85`.
`C(X)`: se define como el número de enlaces que salen de la página `X`.

PageRank establece una distribución de probabilidad sobre las páginas Web, de tal modo que la suma de todos los valores PageRank de las páginas Web será igual a 1.

La idea sobre la que se basa PageRank es bastante intuitiva, asume que si una página recibe bastantes enlaces provenientes de otras, entonces se supone que esa página merece ser visitada. No obstante, también tiene en cuenta el hecho de que páginas muy importantes enlacen a otra, lo que implica que es probable que esa página también sea digna de ser visitada al estar enlazada por una página de calidad.

2.4.2.2 Algoritmo de ranking HITS

El algoritmo *Hypertext Induced Topic Selection* (HITS) se basa en un modelo de la Web que distingue concentradores y autoridades. Como fue mencionado anteriormente, una página es una autoridad en un tema si contiene información valiosa y de alta calidad sobre ese tema. Una página es un concentrador si enlaza a buenas autoridades sobre el tema. Cada página tiene asignado un valor concentrador y un valor autoridad.

- El valor concentrador de la página `X` está en función de los valores de autoridad de las páginas que enlaza `X`.
- El valor autoridad de la página `X` está en función de los valores de concentrador de las páginas que enlazan a `X`.

La puntuación se basa en los siguientes principios:

- La calidad de un concentrador se determina mediante la calidad de las autoridades que le enlazan.
- La calidad de una autoridad se determina mediante la calidad de los concentradores a los que enlaza.

Dado un conjunto de `n` páginas Web, el algoritmo HITS primero constituye una matriz de adyacencia `A` de dimensiones `n×n`, cuyo elemento `(i,j)` es `1` cuando la página `i` enlaza con la página `j`, y `0` en caso contrario. HITS se obtiene mediante el cálculo iterativo de los tres pasos detallado en la Ecuación 6:

Ecuación 6

1. Actualizar las puntuaciones de autoridad de cada página.
$$a^{t+1} = A^T \bullet h^t$$
2. Actualizar las puntuaciones de concentrador de cada página.
$$h^{t+1} = A \bullet a^{t+1}$$
3. Normalizar las puntuaciones autoridad y concentrador.

Donde:
`a`: es el vector con los valores de autoridad.
`h`: es el vector con los valores de concentrador.

Si dos páginas Web distintas p_i y p_j están co-citadas por muchas otras páginas Web p_k (ver Figura 13, a), es probable que estén relacionadas en algún sentido. A su vez, si dos páginas Web distintas p_i y p_j co-referencian varias otras páginas Web p_k implica que p_i y p_j tienen ciertos aspectos en común (ver Figura 13, b).

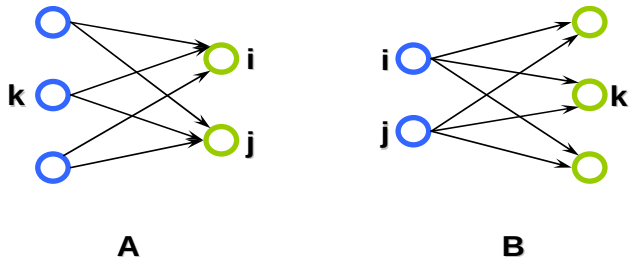


Figura 13: A la izquierda el fenómeno co-citación y a la derecha el fenómeno co-referencia con el algoritmo HITS.

2.4.2.3 Algoritmo de ranking SALSA

El algoritmo *Stochastic Approach for Link Structure Analysis* (SALSA) también asigna dos puntuaciones a cada página: concentrador y autoridad. Estas puntuaciones se basan en dos caminos aleatorios realizados en el grafo G, el camino autoridad y el camino concentrador.

El camino autoridad sugiere que las páginas de mayor autoridad deberían ser enlazadas desde muchas páginas. Así, un camino aleatorio de este sub-grafo visita aquellas páginas con alta probabilidad. Formalmente, el estado del camino autoridad son los nodos de G con al menos un enlace de entrada, los nodos sin enlaces de entrada alcanzan una puntuación 0.

En comunidades en red de alta densidad (*Tightly-Knit Community,* TKC) se destacan las mayores diferencias entre el algoritmo HITS y SALSA (Lempel & Moran, 2000). Una comunidad en una red de alta densidad es un conjunto de páginas pequeño pero sumamente interconectado, en este caso la deficiencia se da cuando los algoritmos de ranking dan a estas colecciones de páginas puntuaciones altas aunque esas páginas no sean autoridades en el tema. HITS favorece a los grupos de páginas que tienen muchas co-citaciones "*internas*", mientras SALSA prefiere las páginas con muchos enlaces de entrada.

2.4.2.4 Ranking personalizado

El método Global KE propuesto por (Akritidis, Katsaros, & Bozanis, 2008) se basa en el algoritmo Original KE (Akritidis, Voutsakelis, Katsaros, Bozanis, & Greeee, 2007). La propuesta plantea incorporar otro tipo de información al ranking que permita que la puntuación se adapte a cada usuario. Global KE considera datos regionales y perfiles de usuario en la puntuación de la relevancia. Además este algoritmo permite el cálculo con uno o múltiples buscadores. Global KE define tres algoritmos:

- El algoritmo *Geo* KE toma en cuenta los datos regionales del usuario y dominio de los resultados a través del coeficiente `G`. Donde `G=2` si la región del usuario es igual al dominio del resultado, `G=3` si el usuario comprende el idioma del resultado, `G=4` si el dominio del resultado no revela información de la localidad del resultado, `G=5` si el usuario no comprende el idioma del resultado.
- El algoritmo *Weighted* KE considera que los buscadores pueden tener pesos o importancia distinta para el usuario. El factor `e(i)` se denomina factor de peso del i-ésimo motor de búsqueda (*Weight Factor of the ith engine,* EWF). Este factor toma valores enteros entre 1 y 10, de más a menos importante.
- El algoritmo UL *aware* KE clasifica los resultados según los dominio/subdominios en la URL de los resultados. Lo anterior se realiza a través de una constante D denominada conciencia del dominio (*domain awareness constant*, DAC), donde `D=5` cuando el dominio del resultado aparece más de 2 veces y `D=10` cuando el dominio del resultado aparece más de 2 veces pero en sub-dominios distintos.

Si bien cada algoritmo puede funcionar de forma independiente, los tres se integran en el algoritmo Global KE el cual es una ampliación del algoritmo original KE. La Ecuación 7 presenta el algoritmo original y la Ecuación 8 presenta el algoritmo Global KE:

Los experimentos realizados con el algoritmo Global KE y el algoritmo original KE no detectaron diferencias importantes. En cuanto a precisión los resultados presentaron mejorías principalmente en aquellas consultas donde el componente geográfico es relevante.

$$W_{ke} = \frac{\sum_{i=1}^{m} r(i)}{n^m * \left(\frac{k}{10} + 1\right)^n}$$ Ecuación 7

Donde:
`r(i)`: es el ranking que ha tomado el ítem `i`.
`n`: representa la cantidad de buscadores que recuperaron el resultado.
`K`: cantidad de ítems incluidos en los ranking.
`M`: total de buscadores.

$$W_{gke} = (11 - D) * G * \frac{\sum_{i=1}^{m} [11 - e(i)] * r(i)}{n^m * \left(\frac{k}{10} + 1\right)^n}$$ Ecuación 8

Donde:
`r(i)`: ranking que ha tomado el ítem `i`.
`e(i)`: se denomina factor de peso del `i`-ésimo motor.
`n`: representa la cantidad de buscadores que recuperaron el resultado.
`k`: cantidad de ítems incluidos en los ranking.
`m`: total de buscadores.
`G`: coeficiente de los datos regionales.
`D`: constante denominada conciencia del dominio.

2.4.2.5 Ranking de relevancia en la recuperación de objetos de aprendizaje

Es común que en la recuperación de OA se utilicen funciones de ranking similares a las utilizadas en la RI en general, básicamente porque en ambos casos se utiliza la búsqueda sintáctica y por lo tanto, el ranking se calcula por la frecuencia de las palabras claves. Esta situación ha cambiado, aunque la mayoría de las búsquedas sigan siendo sintácticas, para evaluar la relevancia que tienen los OA para el usuario es preciso hacer la diferencia y considerar el ajuste a sus preferencias, experiencias, contexto de uso, propósito, entre otros aspectos.

En (Ochoa & Duval, 2008) se propone una función de ranking de relevancia basada en los datos de atención contextualizada extraídos de la interacción entre los usuarios y los objetos de aprendizaje. En esta propuesta se consideran los cuatro tipos independientes de relevancia en la recuperación de información definidos por (Borlund, 2003).

- Relevancia algorítmica. Mide el grado de relación entre la consulta y los resultados.

- Relevancia de tópicos. Mide cuan bien un resultado se encuentra dentro de un tópico dado.
- Relevancia de pertinencia. Mide si el resultado satisface la necesidad de información del usuario, considerando sus preferencias.
- Relevancia situacional. Mide cuan bien el objeto se adapta a la tarea que el usuario esta desempeñando y al contexto en el cual será usado, se incluyen aspectos como el público objetivo, complejidad, entre otros.

Por ejemplo, la relevancia de tópicos puede ser calculada a través de la probabilidad que el OA haya sido descargado, utilizado y re-utilizado en un curso, cuando coinciden las palabras de búsqueda. La relevancia de pertinencia estima las preferencias del usuario a partir de su historial de uso, búsquedas y descargas previas del OA. La relevancia situacional estima el ajuste de los OA con el propósito o con el uso que el usuario quiere darle, con base en el área disciplinaria del curso, rango de edad alumnos del curso, entre otros. La relevancia algorítmica tiene que ver con las funciones tradicionales de similitud de las palabras claves de la búsqueda.

Las funciones que estiman cada una de las relevancias se combinan en una sola función de relevancia. Obviamente que el cálculo de esta función requiere un alto nivel de procesamiento, no obstante los autores plantean que no todos los datos deben ser calculados *on-line*, algunos pueden ser calculados previamente.

2.4.3. Ontologías en la recuperación de información

Las ontologías como modelos de representación de conocimiento han sido incorporadas con distintos propósitos en las etapas del proceso de RI. El potencial de un recurso semántico como las ontologías en el dominio de la RI ha sido demostrado en distintas investigaciones, aunque también ha quedado de manifiesto el alto costo asociado al desarrollo y mantenimiento de las ontologías, lo cual además es proporcional al nivel de detalle y cobertura de la ontología (Croft, 1986).

En un sistema de recuperación basado en ontología, las ontologías también permiten enriquecer semánticamente la indexación de la información, el proceso de formulación y refinamiento de consultas (Stojanovic, Studer, & Stojanovic, 2004) y la evaluación de la relevancia de los resultados (Stojanovic, 2005).

Lo más simple es el uso de ontologías como jerarquía de conceptos en la etapa de formulación de consultas. En este caso es el usuario quien especifica su consulta a través de la navegación en el modelo (L. Huang, 2000; Joho, Sanderson, & Beaulieu, 2004). De esta forma las ontologías dan a los usuarios un punto de referencia para los conceptos y terminología o simplemente sirven como vocabularios controlados.

En la investigación de (Dehors & Zucker, 2006) las ontologías se utilizan para anotar automáticamente los recursos. En la medida que los recursos son etiquetados con base en las ontologías de dominio, es posible enriquecer semánticamente las consultas y

priorizar los resultados de acuerdo al significado y a los conceptos relacionados con la consulta inicial (Morales et al., 2007).

En las investigaciones de (Lee, Tsai, & Wang, 2008; Navigli & Velardi, 2003; Song, Song, Hu, & Allen, 2007; Zou, Zhang, Gan, & Zhang, 2008) se emplean ontologías para extraer nuevos términos que expanden la consulta del usuario. Las expansiones se basan en las relaciones entre los conceptos modelados, principalmente se utilizan relaciones léxicas de sinonimia o términos relacionados.

También se pueden mencionar otras propuestas más especializadas como los repositorios semánticos (Soto et al., 2007) y el espacio inteligente para el aprendizaje SS4L (Dolog et al., 2008).

2.4.4. Expansión de consultas

En la recuperación de información, la expansión de consultas se ocupa del caso en que no exista correspondencia entre la terminología utilizada en la indexación y la utilizada por el usuario para expresar la consulta. Cuando las consultas están bien formuladas, es decir, son menos ambiguas, la posibilidad de obtener buenos resultados son mayores. Por lo tanto, la expansión de consultas se centra en aquellas consultas que no están bien formuladas, son ambiguas o están expresadas utilizando terminología específica a un país o un dominio.

La ambigüedad de una consulta esta directamente relacionada con la cantidad de términos utilizados en ella. Lamentablemente en diversos estudios se muestra la tendencia del usuario a expresar consultas más cortas. Por ejemplo, (Spink, Wolfram, Jansen, & Saracevic, 2001) analizaron tres conjuntos de datos desde más de un millón de consultas enviadas por los usuarios del motor de búsqueda Excite, recogidos en septiembre de 1997, diciembre de 1999 y mayo de 2001. En el año 2001, las consultas contenían una media de 2.6 términos considerando que el 60% contenía sólo 2 términos (Billerbeck & Zobel, 2004).

Para hacer frente a la complejidad en la formulación de las consultas existen distintas estrategias y técnicas de reformulación, refinamiento o expansión de consultas. En términos generales, la expansión de consultas permite hacer búsquedas con variaciones de los términos originales e incluso agregando nuevos términos resultantes de la desambiguación (Bhogal, Macfarlane, & Smith, 2007). Tradicionalmente la expansión se realiza agregando términos a la consulta obtenidos desde los resultados relevantes mejor priorizados de la consulta inicial o a través de algún diccionario o tesauro (Billerbeck & Zobel, 2004).

Una clasificación para las estrategias y técnicas de expansión se da según los mecanismos que éstas aplican, sean interactivos, automáticos o manuales. En las estrategias interactivas, conocidas como retroalimentación de relevancia, el usuario selecciona los resultados relevantes obtenidos de la consulta inicial y a partir de ellos se realiza la expansión. Las estrategias automáticas, tras ejecutar la consulta inicial, asumen que los primeros (n) resultados son relevantes y basados en ellos se extraen los términos

de la expansión. En las estrategias manuales es el mismo usuario quien refina la consulta inicial según los resultados anteriores y realiza una nueva iteración.

El enfoque de extraer nuevos términos desde los documentos mejor priorizados se denomina análisis local. Cuando los términos se extraen desde la colección completa se denomina análisis global. Por último el enfoque mixto combina el análisis global y local. Algunos estudios plantean que como complemento también se deben extraer los términos que no deben aparecer en los documentos resultantes.

Con los métodos interactivos se obtiene mayor precisión aunque se le critica dado que el usuario no siempre esta dispuesto a participar en el proceso (Dunlop, 1997; Ruthven & Lalmas, 2003; Spink, Greisdorf, & Bateman, 1998). Además, se asume que la búsqueda inicial entregará resultados relevantes para que el usuario los seleccione, pero es sabido que dependiendo del conocimiento o experiencia del usuario éste puede errar en la selección, lo que perjudicaría la efectividad de la reformulación de la consulta, a parte de ser un proceso costoso en tiempo y esfuerzo (Abdelali, Cowie, & Soliman, 2007). Con los métodos automáticos la expansión es más sencilla, pero los resultados dependen de que el conjunto de términos relevantes establecidos automáticamente sean de verdad beneficiosos. Cuando la consulta inicial es ambigua, los resultados obtenidos de ella contienen menor cantidad de documentos relevantes. Por lo tanto la extracción automática de términos a partir de estos documentos tampoco suele entregar buenos resultados (Ruthven & Lalmas, 2003).

Como fue mencionado anteriormente, para aumentar la recuperación de documentos relevantes es necesario reducir la ambigüedad de la consulta prestando atención a su contexto. El contexto puede ser tratado de acuerdo a su definición lingüística o basado en la circunstancia. La primera interpreta el contexto como las partes de un discurso que rodean a una palabra o pasaje y que puede revelar su significado. La segunda lo interpreta como condiciones interrelacionadas en las cuales algo existe u ocurre. Claramente, dentro de la RI la primera es la más utilizada (Bhogal et al., 2007). Así también, el contexto puede basarse en un dominio de conocimiento particular o puede estar relacionado con una tarea específica. Existen distintas estrategias para tratar el contexto de la consulta en la RI (Bhogal et al., 2007), por ejemplo:

- La personalización. En este caso el contexto se define según el historial de las consultas y documentos revisados por el usuario. Dicha información se utilizará como base en las futuras búsquedas. La personalización individual se ha ampliado hasta el nivel de comunidad: a través de los perfiles de usuarios se pueden proponer consultas basadas en el historial de usuarios similares.

- El análisis de enlaces. En este caso el contexto se define con base en el contenido y estructura de la información que rodea al texto. Por ejemplo los textos a través de los cuales se realizan los enlaces a la página y la cantidad de enlaces hacia ella.

- Modelos de lenguaje. En este caso los modelos se basan en Lenguajes de Modelado Estadístico (sus siglas en inglés SML). Un modelo de lenguaje es una distribución de probabilidad que captura las regularidades estadísticas de uso del lenguaje natural.

Por lo tanto el contexto se define en función de la frecuencia que dos o más palabras aparezcan próximas.

- Computación ubicua. En este caso se considera que el contexto de la consulta puede ser influenciado por el entorno físico, y que aspectos tales como la tecnología de conexión, comunicación y dispositivos inciden en el contexto de las consultas. Por ejemplo, la consulta hecha por un usuario a través de un teléfono móvil tiene un contexto distinto a la que éste mismo realice desde un computador personal.
- Contexto relacionado con la tarea. En este caso el contexto se establece en función de los términos o palabras que ocurren durante las sesiones de trabajo que realiza el usuario.
- Concepto de estructura de nodos. Las redes léxicas son otra fuente importante para derivar el contexto, pues contienen vocabularios especializados del dominio y relaciones que han sido extraídos automáticamente de una colección de documentos. El vocabulario para la red léxica puede ser desarrollado utilizando herramientas de análisis de texto. Las relaciones léxicas entre los términos son usadas para sugerir términos adicionales.

Las técnicas de expansión de consultas van desde los mecanismos de retroalimentación de relevancia, mencionados antes, hasta aquellos que usan modelos de conocimiento para eliminar la ambigüedad, tales como los tesauros u ontologías (Bhogal et al., 2007).

Algunos métodos incorporados para la expansión de consultas son la co-ocurrencia léxica, el *clustering*, el aprendizaje colaborativo, el *stemming* y los modelos de conocimiento. La co-ocurrencia léxica es el proceso de establecer relaciones entre palabras basado en el análisis de 2 o más términos que están ubicados al lado o próximos en un documento. En este caso los nuevos términos se extraen desde los términos que co-ocurren con el término buscado. En el *clustering* los documentos similares es decir que comparten un número significativo de palabras son agrupados y las palabras representativas de cada cluster son utilizadas para la expansión de la consulta original, es decir, se asume que los documentos similares son relevantes para las mismas consultas. De esta misma forma en el aprendizaje colaborativo se agregan los términos que han sido relevantes en las consultas de los mismos términos realizadas por usuarios "similares". El *stemming*, como fue mencionado anteriormente, es el proceso en el cual las variaciones de los términos son generados por la agregación o remoción de prefijos y sufijos según corresponda. En este caso la expansión puede agregar a la consulta todas las variaciones morfológicas del término buscado o realizar la búsqueda sólo con las raíces de los términos buscados. En la expansión basada en modelos de conocimiento ya sean dependientes o independientes de la colección, los nuevos términos se extraen desde el modelo, es decir, se extraen los términos semánticamente relacionados a la consulta (Bhogal et al., 2007; Chli & De Wilde, 2006).

Un corpus lingüístico es una colección de texto unificado bien estructurado, balanceado y comúnmente anotado (Hilera et al., 2005). En el corpus se establecen las relaciones entre las palabras, lo que permite extraer el sentido que éstas toman. Pues el sentido que las palabras toman en un contexto está definido por las palabras que le acompañan.

A diferencia de los métodos basados en tesauros u ontologías, en la co-ocurrencia léxica y el *clustering* la fuente a partir de la cual se obtienen los términos para hacer la expansión es el conjunto de documentos y no un modelo de conocimiento (Bhogal et al., 2007) .

En (Chu, Liu, & Mao, 2002) se presenta una técnica de expansión de consultas basada en conocimiento dependiente del corpus. El nivel de relevancia de un término específico en la consulta resultante es determinado por su co-ocurrencia con el concepto general asociado al término, el cual es extraído desde el corpus. El concepto general en una consulta es sustituido por un conjunto de términos específicos usados en el corpus que co-ocurren con el concepto clave de la consulta.

La jerarquía de conceptos también puede ser usada para implementar las técnicas de expansión de consultas. (Joho et al., 2004) plantean la búsqueda basada en conceptos en vez de la búsqueda basada en cadenas de texto. Las jerarquías de conceptos son generadas automáticamente desde la colección de documentos. A partir de los documentos priorizados en el tope se extraen las palabras relevantes y se organizan jerárquicamente usando una función de inclusión (*subsumption*) para determinar no sólo que los conceptos están relacionados sino que además la forma cómo se relacionan. Es decir si un concepto está incluido en otro entonces el primero es el padre en la jerarquía (los términos que co-ocurren son agrupados). Para indicar si un término es general o específico se utiliza la medida de frecuencia inversa de documento (*inverse document frequency,* IDF). Los términos ambiguos tienen entradas separadas en la jerarquía, es decir un hijo puede tener más de un padre (grafo dirigido a-cíclico).

En el estudio de (Agirre & Rigau, 1996) se propone un método para resolver la ambigüedad léxica de un texto maximizando la distancia conceptual entre conceptos. La distancia conceptual fue definida por (Rada, Mili, Bicknell, & Blettner, 1989) como "la longitud de la ruta mas corta que conecta los conceptos en una red de jerarquía semántica". Dentro de una ventana de texto se procesan sólo los sustantivos, se comienza por desambiguar el sustantivo que se encuentra en la mitad y los sustantivos contiguos a él se utilizan como referencia o contexto. Para seleccionar el sentido que toma un sustantivo en la ventana de texto se estima la densidad conceptual entre los sentidos o significados (*synset*) extraídos de WordNet y el resto de los sustantivos de la ventana. WordNet es un diccionario que se basa en reglas léxicas establecidas. Los sustantivos, verbos, adverbios y adjetivos están organizados en relaciones semánticas dentro de conjuntos de sentidos (*synset*). Algunos ejemplos de relaciones semánticas usadas son *sinónimo, antónimo, hipónimo* y *merónimo* (Bhogal et al., 2007).

Por otro lado, el algoritmo de recuperación de información basado en la expansión semántica y la clasificación (QEC), propuesto por (Yue, Chen, Lu, Lin, & Liu, 2005) , trabaja sobre la base del modelo de espacio vectorial, la retroalimentación de relevancia propuesta por (Rocchio & Salton, 1971) y el algoritmo de extracción de frases de (Zhong, Chen, Lin, & Yao, 2004). Los algoritmos de extracción de frases parten de la premisa que la intención del usuario puede ser expresada mejor a través de las frases en comparación con el poder expresivo de los términos, asumiendo que una frase es menos ambigua que un término aislado. Primero se clasifican los documentos de la colección, definiendo para cada clase un vector con los **términos clave**. Para cada documento se

aplica el algoritmo de extracción de frases, y a partir de los vectores de frases de los documentos se define el vector de **frases clave** de cada clase. De acuerdo a la pseudo-relevancia, después de aplicar la consulta inicial se extraen los términos para hacer la expansión desde los primeros `m` documentos recuperados en el tope de la lista. Dado que existen 2 consultas; inicial y expandida, se determina la similitud de ambas con los vectores de las clases identificadas. La consulta inicial se analiza con el vector de **frases clave** de cada clase y la consulta expandida con el vector de **términos clave** de cada clase. A partir de esto se determinarán las clases más relevantes. Finalmente los documentos más relevantes se determinan según la similitud entre la consulta inicial y expandida con los vectores de los documentos de la clase más relevante. Las mejoras detectadas con QEC, en cuanto a velocidad y precisión, respecto a los métodos tradicionales se explican por la reducción del espacio de búsqueda dentro de clases.

En general, después de seleccionar los términos para expandir la consulta, se debe considerar el peso de éstos en la relevancia de los resultados. La determinación del peso de los términos expandidos es un factor que incide en los resultados de la expansión. Para asignar el peso a los nuevos términos se han utilizado estrategias tales como darles un peso más bajo que los términos originales, asignarles un peso en función de su ocurrencia en los documentos relevantes, o asignarles un peso según el peso asignado por usuarios similares (filtrado colaborativo), o por omisión darles un mismo peso a todos.

Las consultas cortas ponen a prueba los métodos tradicionales de RI en cuanto a desempeño, precisión y exhaustividad (estas métricas son tratadas en la sección 2.4.2). Cuanto más corta es la consulta mayor es su ambigüedad y por ende, mayor es la posibilidad de recuperar información irrelevante. Si bien la expansión de consultas no es siempre aplicable, en general hay acuerdo en que es apropiada para consultas cortas.

Aún existe mucha investigación y debate respecto a los factores que inciden en la expansión de consultas. Por ejemplo aspectos como la cantidad de documentos que se asumirán como relevantes, la forma de seleccionar los nuevos términos o frases, la cantidad y el peso de los términos expandidos. Según (Bhogal et al., 2007) cada una de estas cuestiones inciden en los resultados y desempeño de la expansión. En general la validez y beneficios de la expansión dependen de las técnicas o los métodos utilizados, la colección, y del tamaño, del tipo o del dominio de las consultas sobre las cuales éstas se aplican.

Hasta hoy, no hay acuerdo en cuanto a la cantidad de nuevos términos. Algunas investigaciones proponen desde dos hasta doscientos o más términos, en lo que sí existe acuerdo es que la calidad de los términos agregados es más importante que el número de ellos (Song et al., 2007).

Como fue mencionado anteriormente, la expansión de consultas no siempre resulta ser beneficiosa, incluso cuando los términos agregados son imprecisos las técnicas de expansión de consultas pueden ir en desmedro del desempeño en la recuperación de información (Abdelali et al., 2007).

En lo relativo a la expansión de consultas, se debe destacar que:

- Las consultas cortas son candidatas para expandir, pues se supone que son más ambiguas.
- La expansión es más exitosa si todos los nuevos términos son relevantes, es decir si con su incorporación mejoran los resultados (Bhogal et al., 2007).
- Según lo han demostrado (Mandala, Tokunaga, & Tanaka, 1999) y (X. Huang, Huang, & Wen, 2005) es recomendable emplear más de una técnica de expansión. Por ejemplo la mezcla de las técnicas de expansión de consultas con ontologías, tesauros y co-ocurrencia ofrece mejores resultados.
- (Kang & Kim, 2003) plantean que según una clasificación de las consultas es posible analizar si una consulta debe o no ser expandida o el método de expansión más apropiado para ella.

2.4.5. Expansión de consultas basadas en modelos de conocimiento

Los tesauros y las ontologías son los modelos de conocimiento más utilizados en la expansión de consultas. Dependiendo del origen del conocimiento representado los modelos pueden ser dependientes o independientes del corpus. Por ejemplo un tesauro creado a partir de los documentos de la colección es *dependiente* de su contenido. En cambio si el modelo es creado por expertos respecto a su visión, conocimiento y experiencia en el dominio estamos frente a un modelo *independiente* del corpus.

Las primeras investigaciones de expansión basadas en modelos de conocimiento se refieren a los tesauros de similitud. Estos modelos son definidos como una estructura de términos similares extraídos de forma automática desde una colección de documentos. Comúnmente para medir la similitud se utiliza la función coseno normalizada o la fórmula `tf·idf` (donde *tf* es la frecuencia del término e *idf* es la frecuencia inversa del documento) (Salton & McGill, 1986).

La expansión de consultas usando modelos de conocimiento dependientes del corpus es más aplicable para colecciones estáticas de documentos, o por lo menos colecciones cuyo contenido tiene poca variación en el tiempo. Para colecciones Web estos modelos deberían ser actualizados o regenerados constantemente debido a la naturaleza dinámica de esta colección, en consecuencia el costo asociado a la actualización de la base de conocimiento desestima su uso (Bhogal et al., 2007).

Por su parte, las ontologías de dominio proveen vocabularios consistentes y representaciones del mundo para la comunicación clara en un dominio de conocimiento. Lo anterior es fundamental considerando que en el lenguaje natural una palabra puede tener múltiples significados dependiendo del contexto donde se aplica. En un sistema computacional el contexto puede ser representado y restringido por una ontología. Las ontologías de dominio específico modelan términos y conceptos que son específicamente usados en un dominio dado.

En general, según (Bhogal et al., 2007) los desafíos a la hora de utilizar modelos de conocimientos para la expansión de consultas se encuentran en:

- Falta de correspondencia entre los conceptos modelados y los términos de la consulta. Para resolver esto es posible realizar la formulación de las consultas utilizando el modelo como un vocabulario controlado o realizar un proceso de correspondencia en el caso que no exista una equivalencia directa entre la consulta y los conceptos modelados.
- Falta de un modelo para un dominio particular. El diseño y construcción de una ontología desde cero requiere mucho esfuerzo, no solo desde el punto de vista técnico sino que de extracción de conocimiento desde los expertos del dominio, de acuerdo y consenso.

2.4.5.1 Revisión de propuestas para la expansión de consultas basadas en modelos de conocimiento dependientes de la colección

A continuación se describen propuestas para la expansión de consultas cuyo elemento común es que plantean el uso de modelos de conocimiento en la expansión de consultas en los sistemas de RI y que dichos modelos son creados a partir de los documentos de una colección, es decir que el contenido del modelo de conocimiento es dependiente de la colección.

- En (Sangoi Pizzato & Strube de Lima, 2003) se evalúa la expansión con tesauros basada en las distintas relaciones léxicas, por ejemplo términos generales, específicos, sinónimos, términos relacionados, entre otros. Las relaciones utilizadas en la expansión son ponderadas por el usuario en función de su importancia percibida. Una vez encontrado el término de la consulta en el tesauro se extraen los términos más cercanos según la medida de distancia semántica. Esta medida pondera el tipo de relación y el peso asignado a los términos relacionados al momento de la creación del tesauro. Una restricción es que los términos seleccionados para la expansión deben ser relevantes para todos los términos de la consulta. La evaluación presentó mejoras en cuanto a exahustividad (*recall*) aunque con reducciones en la precisión. Cabe destacar que la evaluación no se realiza sobre una colección estándar de prueba que permita su comparación con estudios previos.
- La propuesta de (Zazo et al., 2005) para expansión de consultas basado en un tesauro, en este caso se invierte la interpretación tradicional dada a los términos y los documentos en la generación de un tesauro de similitud. Cada término de la colección es caracterizado por los documentos en los cuales aparece. Es decir se invierte el rol de los documentos y términos en el cálculo de la frecuencia inversa de los términos (*idf*). Si 2 términos co-ocurren en un documento, la probabilidad de que puedan ser similares aumenta si el documento es extenso y disminuye si es breve. El tesauro de similitud se construye con todos los pares de similitud calculados para cada uno de los términos de la colección. Los términos seleccionados para la expansión son aquellos que poseen mayor similitud con cada uno de los términos de la consulta, es decir la consulta es tratada como un concepto y no como conjunto de

términos independientes La evaluación entregó buenos resultados de la expansión, especialmente en consultas cortas. La mayor desventaja son los altos costos de construcción y actualización del tesauro lo cual restringe su aplicabilidad a colecciones dinámicas, y descarta su uso en Internet.

- En (Y.-F. Huang & Hsu, 2008) se propone la expansión de consultas a través de un modelo de conocimiento basado en el corpus Pubmed[11], dicho modelo es a su vez complementado con los términos de la Ontología Gene. El árbol de palabras relacionadas se construye sólo con los términos de mayor relevancia de cada *abstract*, denominados *keywords* candidatos. La relevancia de los términos se calcula a través de la fórmula `tf·idf` (frecuencia inversa del documento *idf* y la frecuencia del término *tf*). El árbol se crea por niveles, el primer nivel representa los términos con mayor `tf·idf` en la colección completa, para el segundo nivel se recalcula la frecuencia en función de la colección de *abstract* asociados al nodo que se desea detallar. El método de expansión busca en el árbol los términos de la consulta y se expande a través de los términos *implícitos* que se encuentran en el sub-árbol cuya raíz es el término consultado. Los términos de expansión son seleccionados según el *idf* ponderando además el nivel de profundidad. El experimento desarrollado sólo evalúa la validez de los términos extraídos para la expansión, es decir no evalúa la precisión o exhaustividad de los resultados recuperados con las consultas expandidas.

2.4.5.2 Revisión de propuestas para la expansión de consultas basadas en modelos de conocimiento independientes de la colección

Al contrario de las propuestas descritas en la sección anterior, a continuación se detallan algunas propuestas que plantean el uso de modelos de conocimiento creados independiente a una colección, es decir que el conocimiento representado en el modelo fue extraído de otras fuentes, de la experiencia y conocimientos de los expertos, de la recomendación de los usuarios, desde documentos de referencia en un dominio, etc.

- La propuesta de (Navigli & Velardi, 2003) utiliza la información del *sentido de la consulta* y las ontologías para la expansión. En este estudio se plantea que la expansión con sinónimos e hiperónimos tiene un efecto limitado en el desempeño de la RI en Web y por lo tanto, incluir otro tipo de información semántica puede resultar más efectivo en la expansión. En este caso se propone utilizar la información de la glosa[12] (*gloss*) y las palabras comunes, es decir palabras en el mismo dominio semántico y en el mismo nivel de generalidad. Para cada una de las palabras de la consulta y cada uno de sus sentidos (*synset*) se crea una red semántica, extraída desde WordNet, donde se representan relaciones tales como de ($\rightarrow^{\&}$) sinónimo, ($\rightarrow^{@}$) hiperónimo, ($\rightarrow^{\sim}$) hipónimo, ($\rightarrow^{\#}$) merónimo y ($\rightarrow^{\%}$) homónimo (ver la

[11] http://www.ncbi.nlm.nih.gov/PubMed

[12] Glosa es una nota hecha en los márgenes o entre las líneas de un libro, para explicar el significado del texto, frecuentemente para proporcionar una traducción a otra lengua.

Figura 14). Además se incluye las relaciones de términos que aparecen en la glosa y los tópicos. El método de expansión de nodos comunes intersecta las redes semánticas y selecciona aquellos términos más comunes para finalmente crear un vector de términos comunes ordenados según su ocurrencia. En el experimento se utilizan 24 de las 50 consultas de prueba de la colección TREC2001 las cuales se consultan en Google. Los resultados se evalúan en función de la cantidad de resultados correctos recuperados. Es preciso notar que las mejoras con el método de nodos comunes no son concluyentes ya que dicho método no se aplica a todas las consultas probadas.

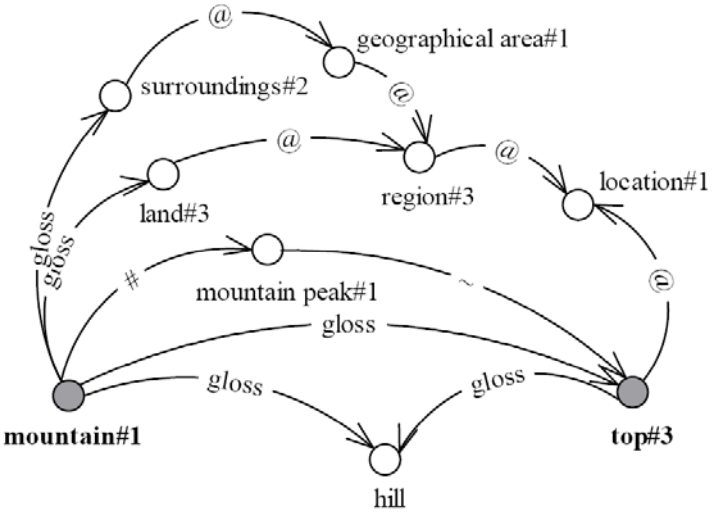


Figura 14: Intersección de la red semántica para la palabra *mountain (synset* 1) y la palabra *top (synset* 3). Los nodos comunes en este ejemplo son *location* y *hill* (ilustración reproducida de (Navigli & Velardi, 2003), pp. 44.)

- En (Sihvonen & Vakkari, 2004b) se analiza si el tesauro *ERIC descriptors* (ERIC, 2010) apoya la expansión de consultas de usuarios tanto expertos como noveles y si además, este aporte se ve reflejado en el éxito de la búsqueda. La premisa es que el tesauro puede complementar el conocimiento del usuario en el tema, aportando terminología apropiada. Como fue estudiado en (Vakkari, 2002) el conocimiento en el tema del usuario influye en la selección de los términos de búsqueda y también afecta la evaluación de la relevancia de los documentos recuperados. El conocimiento en el tema co-varía con la habilidad del usuario para concretar su necesidad de información y, para identificar y seleccionar los términos apropiados (Vakkari, 2002).

  En la formulación de la consulta los usuarios navegan en el tesauro ERIC y acceden a la terminología apropiada según las relaciones de equivalencia (sinónimos) y las relaciones jerárquicas de los términos de la consulta (términos generales y específicos) (Nielsen & Ingwersen, 1999; Sihvonen & Vakkari, 2004a). Las tareas de búsqueda definidas en el experimento fueron dos: una simple y la otra compleja. En la primera se consideran conceptos generales o familiares para cualquier usuario y en la segunda información con mayor especificidad. Los resultados indican que los

usuarios con mayor conocimiento del tema (expertos) utilizan una mayor cantidad de términos del tesauro y también obtienen una mayor cantidad de resultados relevantes. Los tipos de términos más utilizados son los términos relacionados y los sinónimos, aunque en las tareas de búsquedas simples. En las tareas de búsqueda complejas los tipos de términos más útiles resultaron ser hipónimos, es decir los términos que pertenecen al mismo conjunto, género o clase.

- En (Stojanovic et al., 2004) se propone un proceso de refinamiento de las consultas del usuario que parte de la base de un sistema de recuperación basado en ontología. El proceso consta de 3 fases. En la fase de descubrimiento de la ambigüedad se detecta y analizan las posibles ambigüedades en la consulta del usuario. La fase de interpretación analiza las ambigüedades y evalúa su efecto en las metas del usuario, por último en la fase de refinamiento de la consulta se proponen los posibles refinamientos para la consulta, ordenados según su relevancia para que el usuario satisfaga su necesidad de información. Se definen 2 tipos de ambigüedades, la ambigüedad semántica y la ambigüedad relacionada con el contenido. El primer tipo esta determinado por los conceptos a los que pertenece la consulta del usuario y las relaciones que este concepto posee en la ontología. El segundo tipo puede ser medido a través de la comparación de los resultados obtenidos por distintas consultas, si la lista de resultados de una consulta es igual a la lista de otra consulta es un indicador de ambigüedad. El proceso de refinamiento no fue evaluado y sólo se proponen algunas áreas de aplicación.

- En la investigación realizada por (Wollersheim & Rahayu, 2005) se presenta un *framework* para la evaluación de algoritmos de expansión basados en ontologías. Para determinar con que tipos de expansión se obtienen mejores resultados, se extraen los conceptos asociados a los documentos relevantes de la colección de prueba, dichos conceptos son comparados con los términos expandidos de 4 algoritmos. La colección de prueba utilizada en la evaluación es *Ohsumed*. Dicha colección es específica para el área médica, contiene 355.013 documentos (títulos y *abstract*) de MEDLINE entre los años 1987 a 1991 (Hersh, Hickam, Haynes, & McKibbon, 1994). La ontología utilizada para la expansión es *Unified Medical Language Systems*, UMLS. UMLS reúne un conjunto de vocabularios de terminología médica dividido en 2 secciones, meta-tesauro y la red semántica. La mayoría de las relaciones utilizadas en la expansión se extraen desde el tesauro por ejemplo se utilizan las relaciones *Parent, Broader, Child, Narrower, Like-Synonym, Sibling, Relation Other, Accepts Qualifier, Qualified By*. Desde la sección red semántica sólo se extrae el *concepto semántico* relacionado con cada *concepto en el tesauro*. Para la evaluación de los algoritmos, primero se relacionan manualmente las consultas de prueba con el concepto de la ontología más específico asociado a todos los términos de la consulta. Por otro lado, se realiza la asociación de los conceptos extraídos de cada documento relevante con los conceptos de la ontología. Los algoritmos de expansión probados son de profundidad, de probabilidad, de peso semántico y de conceptos interconectados. El primero realiza la expansión directamente siguiendo todas las relaciones del concepto, sin considerar el tipo de relación o la profundidad. El segundo realiza la expansión considerando la probabilidad de cada término

relacionado. La probabilidad disminuye en función de la cantidad de relaciones y del nivel de profundidad. El tercer algoritmo es similar al anterior pero además de la probabilidad considera la importancia de cada tipo de relación (a través de un peso asignado a cada una). El último algoritmo considera que los términos más interconectados son más provechosos en la expansión que los términos aislados. Según la precisión de los resultados del experimento el tipo de relación con mejores resultados es *Broader,* los algoritmos de profundidad y probabilidad obtienen niveles similares de precisión, el algoritmo de pesado semántico obtiene mejores resultados que los dos primeros algoritmos. El algoritmo de conceptos interconectados obtiene mejores resultados que el algoritmo de profundidad aunque en la mayoría de los casos no es aplicable.

- En (Song et al., 2007) se propone la técnica de expansión semántica -SEMANQE. En su algoritmo de expansión se combinan reglas de asociación, ontologías y procesamiento de lenguaje natural. Esta técnica utiliza la semántica explícita, las propiedades lingüísticas del corpus no estructurado, las propiedades contextuales de los términos descubiertos por las reglas de asociación y las ontologías para quitar la ambigüedad. Se incluye el sistema de recuperación *Lemur*, el algoritmo *Apriori*, la técnica de extracción de palabras claves *Gain*, la técnica de etiquetado gramatical (*Grill Pos Tagger*) y se utiliza el diccionario WordNet como una ontología.

  El buscador Lemur funciona de modo *backend*, recibe la consulta inicial del usuario y recupera un ejemplo de documentos. Los documentos recuperados se dividen en sentencias sobre las cuales se aplican las técnicas de procesamiento de lenguaje natural para seleccionar las frases importantes (técnica basada en *Gain y Grill Pos Tagger*). A continuación se aplica un algoritmo de expansión híbrido que combina reglas de asociación y WordNet para generar consultas dirigidas a obtener nuevos documentos similares a los ejemplos. Dado que un término puede tener varios sentidos (*synset*) en WordNet para desambiguar se analiza la similitud entre las frases clave de WordNet y las frases extraídas de los documentos recuperados. Por último, las consultas se reformulan según los resultados del paso anterior.

- En (Ali & Khan, 2008) se propone una solución para la expansión de consultas basados en ontología aplicadas sobre distintas fuentes de datos (estructuradas, semi-estructuradas, o no estructuradas). La expansión se basa las reglas de ampliación de vocabulario y de composición. La primera regla extrae nuevos términos utilizando las relaciones se sinónimos, padre/hijo, variante léxico y acrónimo. La regla de composición extrae términos de la relación parte-de (merónimo/holónimo). La evaluación de la propuesta se realiza sobre la ontología "productos" creada para el estudio. Si bien se concluyen mejoras con la solución propuesta, es preciso mencionar que no se describe el uso de una colección de prueba, y además no se detalla el set de consultas aplicadas o la evaluación de relevancia de los resultados, lo cual resta validez a la propuesta.

- En (Zou et al., 2008) se propone un *framework* basado en ontologías para la expansión semántica de la búsqueda, compuesto por los siguientes elementos:

- Una ontología de dominio construida para el *framework*, denominada CSO *Computer Science Ontology*. En ella se representan los conceptos relacionados con las Ciencias de la Computación. Se modelan tres tipos de relaciones "es parte", "es instancia" y "es tipo".
- El algoritmo de anotación semántica.
- El algoritmo de razonamiento de la expansión semántica.

La ontología CSO (representada en la Figura 15) queda definida como una tupla:

$$DO = \{C, R, H^c, I, A\}$$

Donde:
C: son los conceptos.
R: las relaciones entre conceptos e instancias.
H^c: representa la estructura jerárquica entre los conceptos.
I: las instancias.
A: los axiomas del dominio.

En este *framework* la búsqueda y expansión se realiza sobre recursos previamente etiquetados e indexados a través de la misma ontología utilizada para la expansión de las consultas. La expansión de términos se realiza con los descendientes directos de los conceptos y las instancias. Los resultados de la experimentación son mejores en cuanto a precisión pero son menores en cuanto a la exhaustividad (Zou et al., 2008).

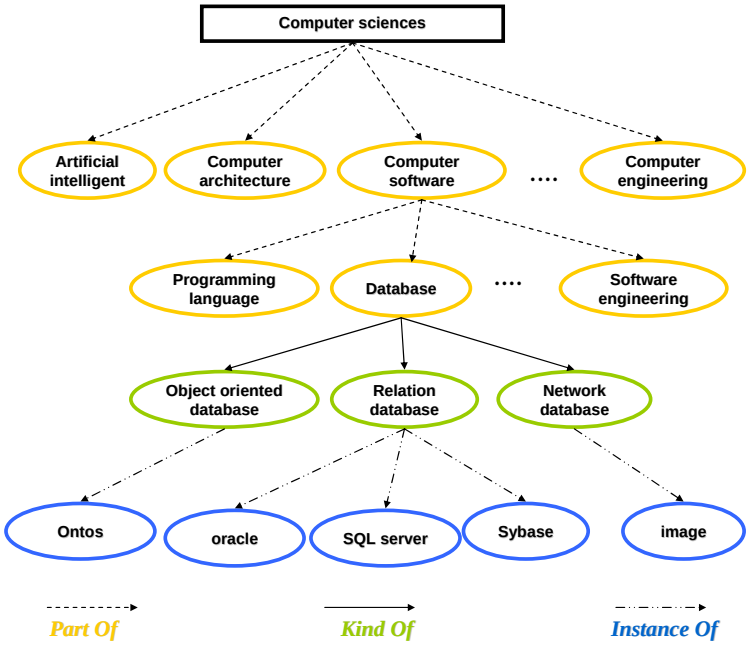


Figura 15: Parte de la ontología CSO. Ilustración reproducida desde (Zou et al., 2008), pp. 454.

- En (Calegari & Pasi, 2008) se propone un método de expansión de consultas que considera las consultas pasadas del usuario y los conceptos relevantes contenidos en los documentos almacenados en su estación de trabajo. Se utiliza una *Object-Fuzzy Concept Network* (O-FCN), es decir una ontología donde los conceptos se relacionan a través de una relación difusa no-taxonómica llamada correlación. La ontología es creada para cada usuario y debe ser dinámicamente actualizada. El FCN se genera del análisis del *stream* de las consultas formuladas por el usuario. La propuesta solo fue probada a través de un ejemplo práctico.

- El modelo de expansión semántica basado en ontología propuesto en (Jiuying Qin, Wang, & Shao, 2009), primero extrae las palabras claves que representan la consulta y se expanden a través de las relaciones padre/hijo y sinónimos. Las palabras candidatas con mayor similitud y relatividad con las palabras de la consulta del usuario son seleccionadas para la expansión. La evaluación del modelo se realiza sobre una ontología creada para este propósito, lamentablemente no existe detalle de la colección de prueba, y tampoco de las consultas y la evaluación de relevancia utilizada.

- En (Ma, Chen, Gao, & Yang, 2009) se propone la expansión de consultas a través del conocimiento disponible en distintas fuentes, tales como tesauros, lexicón, libros de nombres propios o vocabularios (*onomasticon*), base de datos de hechos y ontologías. El analizador semántico extrae de la consulta los pares de atributos que reflejan una expresión regular, por ejemplo de la forma "*1000 US, computer→ precio*: *1000 US*". Luego sobre la base del *onomasticon* se extraen de la consulta las entidades/sustantivos. Los términos restantes son considerados la consulta inicial. Para cada uno de los términos de la consulta se construye un diágrafo semántico, similar a la propuesta de (Navigli & Velardi, 2003). Un diágrafo semántico describe la relación semántica entre las palabras y frases (representados como vértices), donde las aristas etiquetan el tipo de relación (merónimo, holónimo, hiperónimo, hipónimo, sinónimo, etc.) y su peso. Finalmente, son seleccionados los vértices cuya distancia semántica al vértice inicial en cada diágrafo es menor que el umbral. La evaluación de la propuesta se realiza sobre una ontología creada para el estudio en el dominio de turismo. Aunque no existen detalles de las evaluaciones de relevancia realizadas a los resultados con y sin expansión, los autores indican que se obtuvieron mejoras significativas.

- En la investigación de (Lee et al., 2008) se propone un algoritmo para realizar la expansión de las consultas con base en una ontología de dominio específica *JAVA Learning Object Ontology-JLOO* propuesta en (Lee, Yen Ye, & Wang, 2005). El marco de trabajo propuesto se divide dos partes:

    - Capa de aplicación (*backend*). En este modo se encuentran los repositorios de OA. Los metadatos de cada OA son indexados a uno de los conceptos de la ontología.

    - Capa de interfaz (*frontend*). En este modo se recibe la consulta del usuario y se extraen las palabras claves. Luego se determina la intención del usuario

representada como un sub-conjunto de la ontología (sub-árbol) que contiene los conceptos de la consulta del usuario (refiérase a la Figura 16). Para determinar la intención del usuario se calcula el impacto total de los conceptos base y sus conceptos relacionados semánticamente hacia la raíz del sub-árbol. Los términos que serán agregados pertenecen al sub-árbol de la intención del usuario con mayor impacto que superan el umbral.

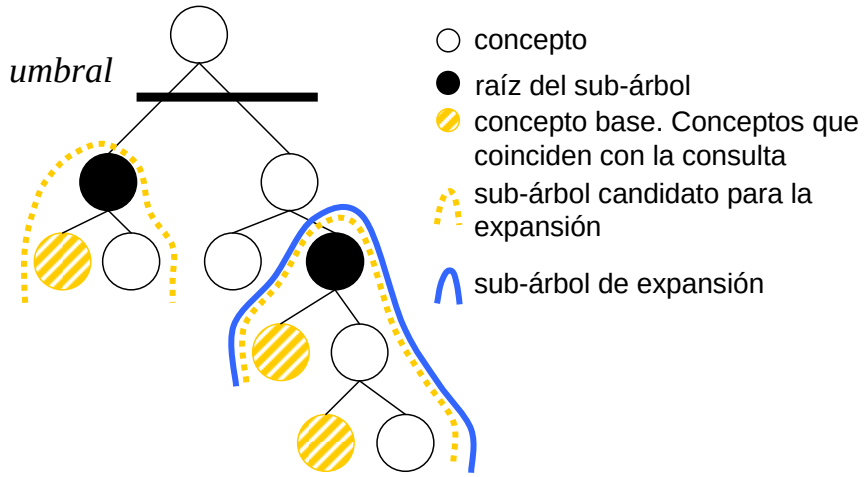


Figura 16: Árbol de intención del usuario, según algoritmo Lee et al. (2008).

Si bien la propuesta de Lee at al. (2008) aborda puntos similares a nuestra investigación, en la sección 2.5 se detallan las características de éste trabajo que representan diferencias significativas con nuestra investigación.

- La investigación presentada por Díaz-Galiano et al. (2009) evalúa el uso de la expansión de consultas en el área biomédica, basada en el tesauro MeSH (Nelson, Johnston, & Humphreys, 2001). La búsqueda se realiza sobre una colección de datos textuales e imágenes. Por lo tanto, se complementan dos sistemas de recuperación; *Lemur* para la información textual y *GIFT system* para la información visual. Aunque MeSH es llamado una ontología, por definición es un tesauro, pues no pasa de ser un vocabulario controlado compilado por la *National Library of Medicine* (NLM) que contiene conceptos, sinónimos cercanos y conceptos cercanos relacionados.

  Para la expansión, primero se realiza la correspondencia entre la consulta y los términos del tesauro, si existe correspondencia se agregan a la consulta todos los términos sinónimos (en los experimentos estos términos se redujeron sólo a 3 categorías). Cabe mencionar que en la correspondencia (*matching)* entre la consulta y los conceptos del tesauro no importa el orden de las palabras, sino únicamente que estén contenidas en la consulta (Díaz-Galiano, Martín-Valdivia, & Ureña-López, 2009). En la evaluación se compara la efectividad de la expansión en tres situaciones, usando sólo la información textual, la información visual o su combinación. Para la combinación se utiliza un sistema de fusión donde se unen los resultados del sistema de recuperación textual con los resultados del sistema de recuperación visual. En el

caso de resultados repetidos, el valor de *ranking* se re-calcula asumiendo un "efecto coro", es decir que los resultados recuperados por ambos sistemas son mejor ponderados que los recuperados por un sólo sistema. Según su evaluación la expansión mejora los resultados de la recuperación de información textual y la combinación de ésta con la información visual.

- En (Tuominen, Kauppinen, Viljanen, & Hyonen, 2009) se propone la expansión de consultas basada en una ontología de dominio y en una ontología espacio-temporal, dicha propuesta forma parte del proyecto *Finnish Ontology Library Service -ONKI*[13]. La expansión con la ontología de dominio se realiza a través de las relaciones *is_a* y sinónimos, y la expansión con la segunda ontología se realiza a través de relaciones propias tales como *near*, *far*, etc. Las ontologías para la expansión son también las utilizadas en el etiquetado de los recursos del *Collection of Finnish Museums of Forestry*. En cuanto a la evaluación de la propuesta, se debe mencionar que los resultados de la relevancia para el usuario no han sido evaluados.

2.5. Caracterización de la investigación

Una vez realizada la revisión de las principales propuestas de expansión de consultas basadas en ontologías, es necesario destacar los aspectos que hacen la diferencia con nuestra investigación.

En la Tabla 8 se resumen las características de las propuestas revisadas con mayor relevancia para nuestra investigación. Los criterios de análisis son:

1. Ámbito de aplicabilidad. Con este criterio se analizan las restricciones de aplicabilidad de la propuesta en términos del sistema de recuperación de información para el cual fueron definidas.
2. Tipo de modelo de conocimiento. Si bien las propuestas incluidas en la tabla indican el uso de ontologías como modelos de conocimiento, muchas veces esto no concuerda con las características del modelo utilizado.
3. Dependencia del corpus/colección. Con este criterio se especifica si el modelo de conocimiento utilizado en la expansión de consultas es dependiente o independiente del corpus/colección.
4. Origen del Modelo. A través de este criterio se aclara si el modelo de conocimiento utilizado es independiente de la propuesta de expansión o forma parte de la misma.
5. Tipos de relaciones utilizadas en la expansión. Se detallan las relaciones a partir de las cuales se extraen los nuevos términos para la consulta.

[13] http://www.yso.fi, consultado 07 de mayo de 2010.

6. Evaluación de resultados. Se describen los aspectos más importantes relacionados con la colección de prueba, forma de evaluación de relevancia o el conjunto de datos de prueba.

Tabla 8 : Síntesis de las características de las propuestas revisadas con mayor relevancia para nuestra investigación.

Propuesta	Ámbito	Tipo de modelo	Dependencia del corpus	Origen	Tipos de relaciones	Evaluación de resultados
(Navigli & Velardi, 2003)	WWW	Diccionario de propósito general (Wordnet)	Independiente	Externo	Léxicas (sinónimo, merónimo, hiperónimo, holónimo, hipónimo)	El conjunto de prueba utiliza 24 de 50 consultas de *TREC 2001 web track*. La evaluación se realiza en función de la cantidad de resultados correctos.
(Song et al., 2007)	WWW	Diccionario de propósito general (Wordnet)	Independiente	Externo	Léxicas (sinónimo)	El conjunto de prueba es *TREC ad-hoc queries (5, 6, y 7)*.
(Zou et al., 2008)	Búsqueda en un repositorio propio con recursos etiquetados.	Ontología de propósito específico (*computer science*)	Independiente	Creada para la propuesta	Padre/Hijo	Evaluación sobre la colección de recursos indexados. No se conocen detalles sobre la evaluación de la relevancia.
(Calegari & Pasi, 2008)	WWW	Ontología personalizada (por usuario)	Dependiente historial de búsquedas del usuario	Creada para la propuesta	Correlación (relación difusa)	No se realiza una evaluación formal. Sólo se plantea un ejemplo práctico.
(Ali & Khan, 2008)	WWW	Ontología de propósito específico (productos de venta)	Independiente	Creada para la propuesta	Léxicas (sinónimo, acrónimo, variante léxico) Padre/Hijo	No se conocen detalles respecto al conjunto de prueba y la evaluación de la relevancia de los resultados con y sin expansión.
(Jiuying Qin et al., 2009)	WWW	Ontología	Independiente	Creada para la propuesta	Padre/Hijo Hermano	No se conocen detalles respecto a la evaluación de la relevancia de los resultados con y sin expansión.
(Ma et al., 2009)	WWW	Ontología de propósito específico (turismo)	Independiente	Creada para la propuesta	Léxicas (sinónimo, merónimo, hiperónimo, holónimo, hipónimo)	No se conocen detalles respecto a la evaluación de la relevancia de los resultados con y sin expansión.
(Lee et al., 2008)	Búsqueda en repositorio propio con recursos etiquetados e indexados.	Ontología (*java object*)	Independiente	Específica para la propuesta	Léxicas (sinónimo) Padre/Hijo	No se conocen detalles respecto a la evaluación de la relevancia de los resultados con y sin expansión.
(Díaz-Galiano et al., 2009)	Sistema de recuperación de información médica multimodal.	Tesauro	Independiente	Externo	Léxicas (sinónimo)	El conjunto de datos es imageClef 2005 y 2006.
(Tuominen et al., 2009)	Búsqueda en librería propia con recursos etiquetados e indexados.	Ontología de dominio y una ontología espacio-temporal.	Independiente	Externo (dentro del proyecto ONKI).	Léxicas (sinónimo) Padre/Hijo Cercanía (ontología espacio -temporal).	No se conocen detalles respecto al conjunto de prueba y la evaluación de la relevancia de los resultados con y sin expansión.

Bajo el mismo análisis anterior, nuestra propuesta quedaría descrita como:

Propuesta	Ámbito	Tipo de modelo	Dependencia del corpus	Origen	Tipos de relaciones	Evaluación de resultados
	Repositorios de OA (no específico para la propuesta)	Ontología formal de dominio	Independiente	Externo (reconocido y validado por los expertos en el dominio)	Ontológicas básicas y específicas del dominio Léxicas	La evaluación formal.

Una de las propuestas que presenta mayor similitud con nuestra investigación es la planteada por (Lee et al., 2008). Si bien ya ha sido descrita en la sección anterior (2.4.5.2), a continuación se analizan más detalladamente las características que representan diferencias significativas con nuestra investigación.

- La expansión es realizada a través de relaciones terminológicas, las cuales son típicas en tesauros más que en modelos ontológicos. Una vez que los términos de la consulta son encontrados en la ontología, la expansión se realiza con los sinónimos o términos relacionados a los términos de la consulta.

- Tal como fue evaluado por (Segura N., Vidal C., & Prieto M., 2010) este algoritmo no puede ser utilizado con otras ontologías de dominio. Lo anterior se debe a la dependencia con la estructura de la ontología JLOO, es decir las relaciones de sinonimia y los niveles jerárquicos o profundidad de la ontología utilizada. En ontologías con poca profundidad o más "achatadas", es decir una ontología que posee pocos niveles de profundidad jerárquica en la representación de conceptos, la expansión se vería perjudicada. La situación anterior ocurre porque para determinar el sub-árbol de la ontología desde el cual se extraen los nuevos términos, el algoritmo considera el número de conceptos de la ontología que coinciden con los términos de la consulta y el peso de ellos propagado hacia la raíz. El sub-árbol con un mayor peso y que además supera el umbral es utilizado en la expansión. Como se representa en la Figura 17, los conceptos de la ontología a, b y c coinciden con alguno de los términos de la consulta. A la izquierda de la figura, si el impacto de los conceptos del subárbol 2 (s-t2) es mayor al del subárbol 1 (s-t1) y además supera el umbral, la expansión se realiza con todos conceptos del subárbol 2. A la derecha de la figura, en una ontología menos profunda si el impacto de los conceptos en el subárbol 1, 2 y 3 (s-t1,2,3) hacia la raíz superan el umbral la expansión se realizará agregando todos los conceptos de la ontología o bien, si el impacto no supera el umbral no habrá expansión.

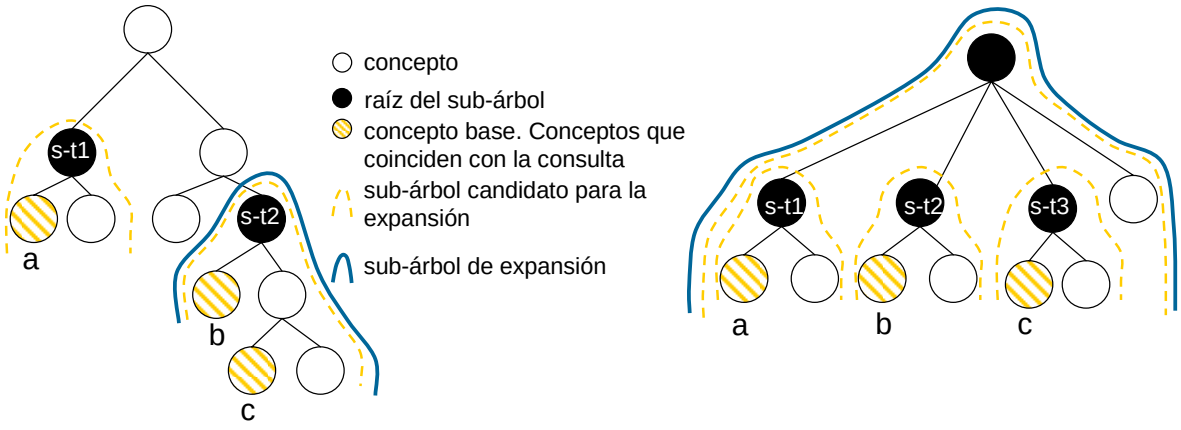


Figura 17: Efecto de la profundidad de la ontología en el algoritmo de expansión propuesto por Lee et. al. (2008).

- De acuerdo a la información disponible de la ontología JLOO, esta fue creada a partir de una colección de documentos y no contiene el conocimiento consensuado de los expertos en el área; tampoco existe una entidad responsable de su actualización.
- En la ontología JLOO no existe un uso consistente de las relaciones de *is_a* (tipo/subtipo). Las relaciones que se encuentran modeladas no necesariamente reflejan relaciones de jerarquía padre/hijo entre los conceptos. Por ejemplo, en la ontología se definen las siguientes clases; "*class declaration*", "*modifiers*", y "*abstraction class*". Estos 3 conceptos son definidos como un tipo de la clase "*class*", es decir un subtipo de esta última. Dichas relaciones no son las más apropiadas en el dominio de la programación en JAVA, por ejemplo el concepto "*modifiers*" puede estar relacionado con el concepto "*class*" a través de una relación "*declarado como*".
- El algoritmo procesa la cadena de búsqueda como un conjunto de palabras, es decir cada una es tratada como un término de búsqueda individual. Por ejemplo, si la cadena de búsqueda es "*method declaration*" los términos que serán buscados en la ontología son "*method*" y "*declaration*" por separado.
- Finalmente, no se utiliza una colección de prueba estándar para la evaluación, no existe detalle respecto a la evaluación de la relevancia de los documentos de la colección y de los resultados de las expansiones.

Capítulo 3.Planteamiento del problema

En este capítulo se describen los problemas identificados después del análisis del estado del arte en el contexto de la búsqueda de objetos de aprendizaje en repositorios y de la expansión de consultas basadas en ontologías. La delimitación del ámbito del problema y las características de nuestra investigación permiten exponer las diferencias entre nuestra propuesta y las propuestas realizadas previamente en otros contextos.

3.1. Problemas generados por el uso de ontologías en la expansión de consultas

El uso de ontologías en la expansión de consultas implícitamente genera algunos problemas o limitaciones importantes de mencionar. A continuación se describen aquellos que son más relevantes para nuestra investigación (Bhogal et al., 2007).

- Falta de correspondencia entre los conceptos de la ontología y los términos de la consulta.

  Cuando la ontología se utiliza en la formulación de la consultas como un vocabulario controlado este problema puede ser irrelevante. Por el contrario cuando el usuario formula libremente la consulta es posible que no exista una equivalencia directa entre la cadena de búsqueda y los conceptos modelados.

- Falta de familiaridad con el modelo de conocimiento.

  Al igual que en el caso anterior, cuando la ontología se utiliza como un vocabulario controlado, la navegación que el usuario realiza en un modelo conocido resulta ser más natural y por lo tanto, aumenta las oportunidades de éxito de la búsqueda. Por otro lado, aún cuando el usuario no interactúe con la ontología la familiaridad de éste con el modelo puede reducir la ambigüedad de la consulta y los problemas de correspondencia descritos anteriormente.

- Dificultad para navegar en el modelo de conocimiento.

Así como en el caso anterior, la facilidad para que el usuario pueda navegar en una ontología aumenta las oportunidades de éxito de la búsqueda. Cuando el usuario no interactúa con la ontología la facilidad de navegación puede ser irrelevante.

- Falta de una ontología para un dominio particular y los altos costos de tiempo y esfuerzo requeridos en su construcción.

  El uso de ontologías en la expansión de consultas supone la existencia y disponibilidad de una ontología de dominio. Si bien hoy en día existen ontologías publicadas, no existen para cada dominio, no fueron creadas a partir del conocimiento experto, o simplemente no han sido validadas por la comunidad de expertos en el dominio. Construir una ontología requiere mucho esfuerzo, no sólo desde el punto de vista técnico sino que, de extracción de conocimiento desde los expertos del dominio, de acuerdo y consenso.

- Baja calidad del modelo de conocimiento.

  La compleción, consistencia, coherencia y validez del conocimiento modelado en una ontología afecta directamente los resultados de la expansión.

3.2. Problemas detectados en la búsqueda de objetos de aprendizaje en repositorios

Después del análisis del estado del arte en cuanto a la recuperación de OA en repositorios, a continuación se resumen algunos de los problemas detectados en el ámbito de esta investigación. Estos problemas son analizados según la dimensión de la búsqueda ya sea respecto al tópico (o tema) al que se refiere el OA buscado, o bien respecto a las restricciones de contexto y audiencia para la cual el OA será utilizado.

- La búsqueda de recursos de aprendizaje según el tópico al que se refiere el OA buscado.

  1. La búsqueda según el tópico tradicionalmente se realiza a través de una búsqueda aproximada dentro de cualquier elemento textual del recurso de aprendizaje, ya sea en su contenido o en cualquier campo de sus metadatos. Lo anterior trae consigo dificultades similares a las detectadas y estudiadas en el campo de la RI en general, como lo es la ambigüedad del lenguaje natural con el que se describe el contenido del OA, sus metadatos o con el cual se especifica la consulta. Cuando la búsqueda aproximada amplía el ámbito de la búsqueda a cualquier elemento textual aumenta la cantidad y la diversidad de los resultados sin que con ello se alcancen mejores niveles de relevancia. Algunas mejoras se presentan cuando se combina con una búsqueda por metadato, permitiendo al usuario especificar a qué campos se limita la consulta por tópico, por ejemplo los campos de título, palabras claves o descripción.
  2. En la búsqueda según el tópico los mecanismos provistos en los repositorios para el refinamiento o expansión de las consultas en su mayoría son manuales puesto

que el usuario refina la consulta a través de la navegación en un tesauro, ontología, jerarquías de conceptos, palabras claves, tipo de recursos, etc. La Figura 18, Figura 19, Figura 20 y Figura 21 ejemplifican algunos casos de refinamiento de consultas en repositorios a la fecha[14].

La Figura 18 representa el mecanismo de formulación y refinamiento de consultas realizado a través de la navegación en un tesauro (ETB-LRE MEC-CCAA V1.0). El usuario navega jerárquicamente a través de los conceptos incluidos en el tesauro, en este ejemplo se presentan los conceptos bajo el concepto general (TG): CULTURA.

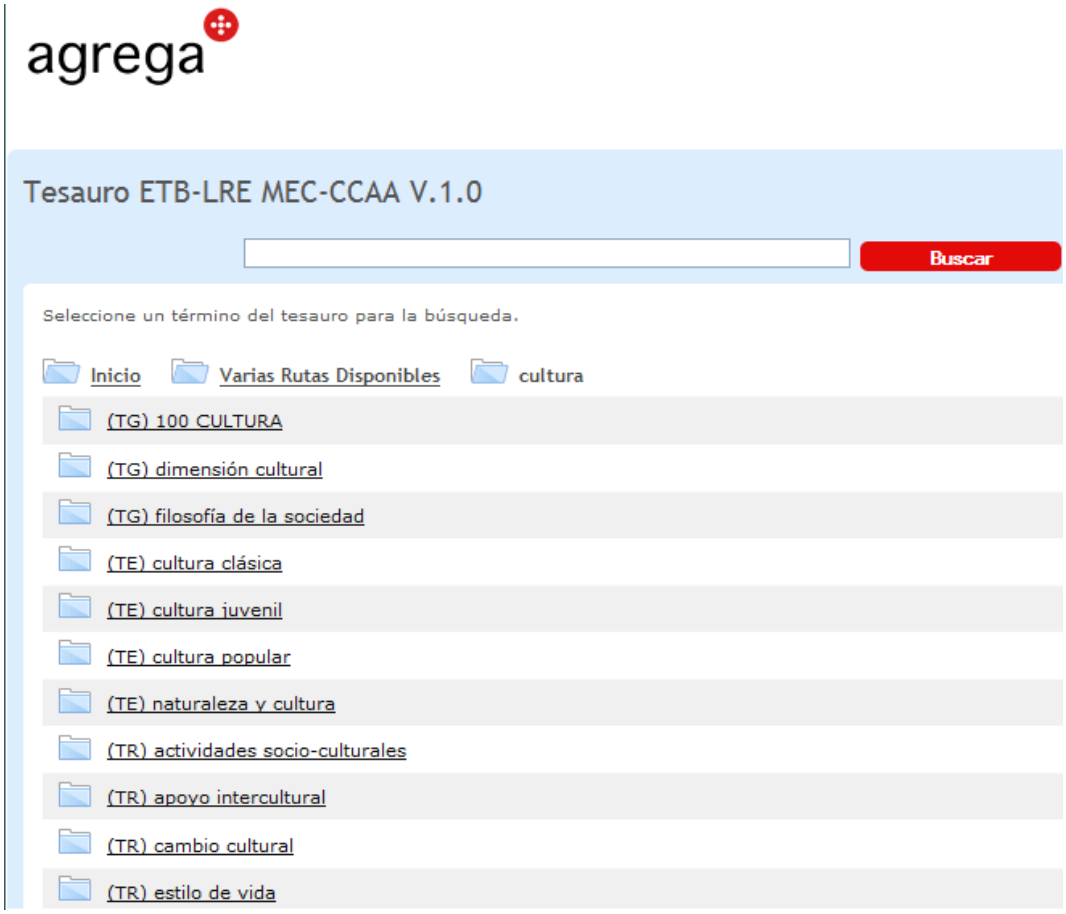


Figura 18: Ejemplo de opciones para la formulación de consultas a través de la navegación en Tesauro ETB-LRE MEC-CCAA, en el repositorio AGREGA. http://www.proyectoagrega.es/default/Inicio.

La Figura 19 y la Figura 20 representan el uso de filtros para el refinamiento manual de consultas, tales como el filtrado por tema, tipo de recurso, nivel educativo, palabras claves, formato, catálogo, autores, etc.

[14] Consultadas realizadas en Abril 2010.

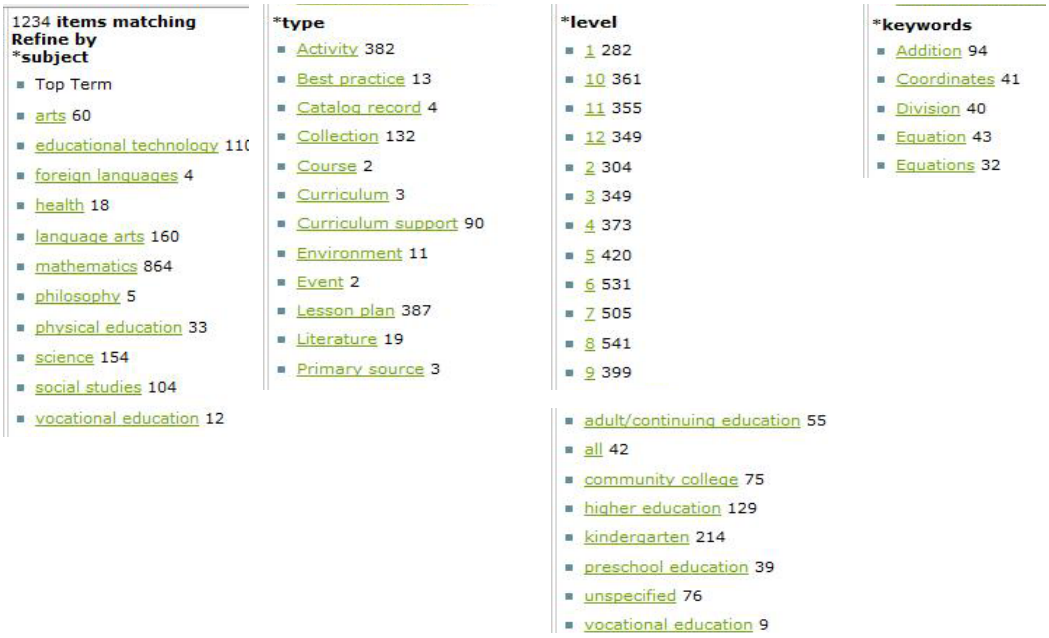


Figura 19: Ejemplo de opciones de refinamiento manual de consultas en el repositorio Gateway.

Figura 20: Ejemplo de opciones de refinamiento manual de consultas en el repositorio Laclo.

La Figura 21 representa el mecanismo para la formulación de consultas a través de la navegación en una ontología de dominio (OA-AE). En la medida que el usuario selecciona un concepto en la ontología (*method* → *agricultural method)* el concepto se expande y se visualizan conceptos relacionados (*agricultural method* → *companion planting,*

monoculture, polyculture, intercropping, soiless culture, etc.). Cada vez que el usuario selecciona un concepto se ejecuta la búsqueda de los OA relacionados con él.

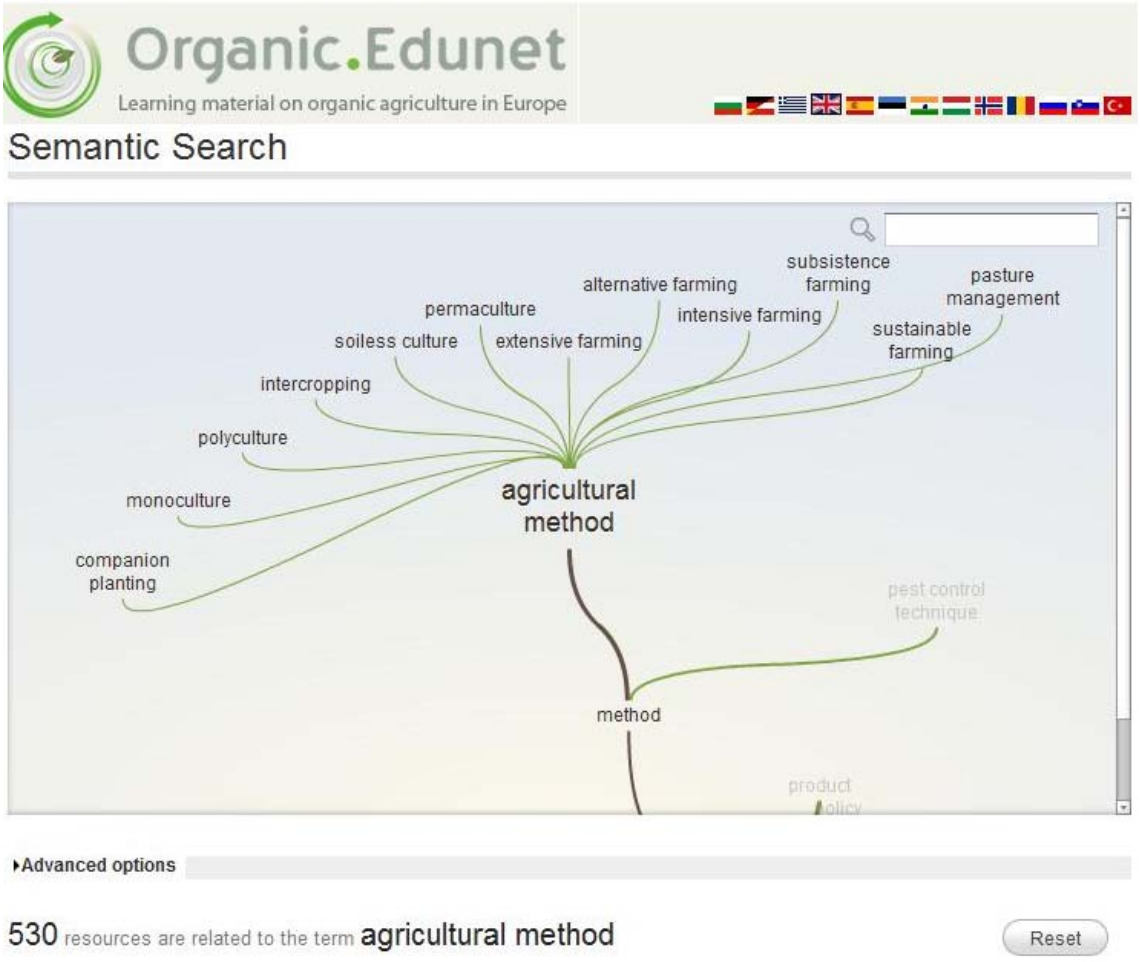


Figura 21: Ejemplo de la formulación de consultas a través de la navegación en una ontología de dominio (OA-AE), en el repositorio Organic.

Algunos de los factores críticos en los mecanismos manuales e interactivos utilizados para la formulación, refinamiento o expansión de consultas es el conocimiento, el tiempo y la participación del usuario, situación para la cual éste no siempre esta dispuesto (Dunlop, 1997; Ruthven & Lalmas, 2003; Spink et al., 1998). De ahí, la importancia de incorporar mecanismos automáticos de expansión de consultas.

- Respecto a las restricciones de la búsqueda según el contexto y la audiencia para la cual el OA será utilizado.

3. Por lo general, para la búsqueda en repositorios sólo se consideran algunos campos de metadatos como filtro, por ejemplo el nivel educativo, el tipo o el formato de los recursos. Por otro lado, las búsquedas que permiten el filtrado según los valores de sus metadatos no siempre utilizan vocabularios estándares. Muchas veces esos vocabularios no concuerdan con el estándar de metadatos o el perfil de aplicación utilizado. Esta situación aumenta la ambigüedad para un

usuario que accede a múltiples repositorios y también reduce la interoperabilidad entre repositorios haciendo más compleja la trazabilidad entre estándares.

En la Tabla 9 se ejemplifican los valores posibles para el campo *nivel educativo* en los repositorios Edna y Merlot, y el campo *tipo de recursos* en los repositorios Intute y Edna.

Tabla 9: Diferencias en los vocabularios y espacios de valores utilizados en los repositorios para los campos nivel educativo y tipo de recurso.

Nivel educativo	
Edna (http://www.edna.edu.au/edna/go)	**Merlot (http://www.merlot.org/merlot/index.htm)**
	Grade School
Adult and Community Education	*Middle School*
Early Childhood Education	*High School*
Higher Education	*College General Ed*
International Education	*College Lower Division*
School Education	*College Upper Division*
Vocational Education and Training	*Graduate school*
	Professional

Tipo de recursos		
Intute (http://www.intute.ac.uk/)		**Edna (http://www.edna.edu.au/edna/go)**
Archives	*Events*	
Arts Projects	*Exhibitions and Galleries*	*Audio*
Associations	*Field Guides*	*Image*
Blogs	*Government publications*	*Interactive*
Case studies	*Images*	*Text*
Collections	*Interactive resources*	*Video*
E-books	…	

4. Con respecto a las restricciones de búsqueda, los repositorios proveen pocas opciones al usuario para especificar las características pedagógicas de los OA, por ejemplo a través de metadatos como el *nivel de interactividad*, la *densidad semántica*, *tipo de interacción*, *dificultad*, entre otros. Aún en el caso de ofrecer estas opciones no proveen asistencia al usuario para especificarlas. En cierta medida se asume que el usuario posee conocimientos especializados en cuanto a los aspectos pedagógicos y al llenado de los campos del estándar de metadato.

Una situación ideal permitiría que el usuario no requiera ser un experto en un estándar para que pueda restringir la búsqueda a un enunciado como el siguiente, por ejemplo, "recursos que permitan el aprendizaje activo para alumnos adultos-mayores principiantes". Una restricción de este tipo, tras el análisis del conocimiento experto, implícitamente se puede traducir en:

```
( (nivel_interactividad="alto" OR
nivel interactividad="medio") AND (tipo de interacción
="activo" OR tipo de interacción ="mixto") AND
densidad semántica="baja" AND dificultad="fácil"… ).
```

5. Referente a las limitaciones en los servicios de búsqueda provistos por los repositorios.

Gran parte de los repositorios además de la búsqueda local permiten la búsqueda en una federación o dentro de una comunidad. Esto provee beneficios en cuanto a la reducción del esfuerzo de ejecutar la búsqueda en múltiples repositorios y a facilitar el acceso a recursos distribuidos. Además de los beneficios este sistema de búsqueda impone restricciones de acuerdo a los estándares de interoperabilidad utilizados. A pesar de las alternativas, tanto para la búsqueda federada como para los servicios externos de consulta, los repositorios utilizan el tipo de búsqueda más simple es decir aproximada conjuntiva en cualquier elemento textual.

Es común que los repositorios impongan ciertas restricciones en cuanto a la forma de expresar las consultas, ya sean que estén determinadas por un estándar de interoperabilidad o por el repositorio en si. Estas restricciones se refieren a aspectos tales como el uso de conectores lógicos, la cantidad de conceptos incluidos en la cadena de búsqueda o simplemente, el lenguaje de consulta.

Para esta investigación es de interés analizar el acceso externo a los repositorios, en dicho caso se han detectado dificultades para invocar el servicio de búsqueda con una cadena que contenga `n` conceptos (`n>1`) y donde cada uno se componga de `m` palabras (`m>1`), tal como es representado en el ejemplo:

```
"concepto buscado 1" OR/AND "concepto buscado 2" OR/AND
"concepto buscado 3" OR/AND...OR/AND "concepto buscado n"
```

En el caso de los servicios bajo el estándar SQI, la mayoría de los repositorios utiliza el lenguaje VSQL, por lo tanto queda descartado el uso de conectores lógicos e incluso existen problemas para tratar un concepto que contiene más de una palabra (el detalle de las características más relevantes de este lenguaje están descritas en la sección 2.2.5).

Con otros tipos de servicios externos de consulta las dificultades son similares. Por ejemplo el servicio RESTful de consulta provisto por el repositorio MERLOT incluye opciones para restringir la búsqueda con los siguientes parámetros *anyKeyWords*, *allKeyWords* o *exactPhraseKeyWords*. Incluso con estas opciones no sería posible tratar una cadena de búsqueda como la ejemplificada anteriormente.

Las restricciones mencionadas son particularmente importantes en la expansión de consultas ya que la consulta inicial se verá extendida con la disyunción de los conceptos expandidos.

3.3. Problemas en la expansión de consultas basadas en modelos de conocimiento aplicadas en repositorios de objetos de aprendizaje

Esta investigación se enfoca específicamente en el uso de mecanismos automáticos para la expansión de consultas basados en ontologías aplicados a la búsqueda en repositorios de objetos de aprendizaje. De acuerdo a los antecedentes recopilados, gran parte de las

investigaciones en esta área están dirigidas a la búsqueda en la Web y no a su aplicación en repositorios especializados de OA. Si bien estas propuestas no son directamente extrapolables, a continuación se sintetizan los aspectos más relevantes de la expansión de consultas basada en modelos de conocimiento, en general y en particular basado en ontologías, aplicadas a la búsqueda en Web, librerías digitales o repositorios institucionales:

- **Expansión de términos basados en lexicón o diccionarios tal como WordNet.**

  En este caso, la expansión se realiza principalmente sobre términos sinónimos, hiperónimo/hipónimo ó merónimo/homónimo. Este tipo de expansión ha sido efectiva para consultas generales, pero sus resultados dejan de ser beneficiosos para consultas específicas en un dominio, ya que existen términos propios o válidos en un dominio en particular o bien términos cuyo significado es distinto en otro dominio.

  Dado que en la recuperación de OA el diseñador instruccional requiere recursos digitales en un dominio específico la utilidad de los términos expandidos con un diccionario general son mínimos. Algunos ejemplos de propuestas en este ámbito, descritas en el capítulo anterior, son (Navigli & Velardi, 2003; Song et al., 2007).

- **Expansión de términos basados en modelos de conocimiento dependientes del corpus.**

  El caso más conocido es el de los tesauros de similitud creados a partir de la frecuencia y co-ocurrencia de los términos extraídos de los documentos de la colección. Algunos ejemplos de propuestas en este ámbito son (Salton & McGill, 1986; Zazo et al., 2005). En esta categoría también se encuentra la expansión basada en taxonomías o jerarquías de conceptos, como por ejemplo en la propuesta de (Zhang, Du, Li, & Jia, 2009) la expansión se basa en la jerarquía de conceptos extraída de la página de directorio de Google.

  El uso de modelos dependientes del corpus impone altos costos de actualización en la medida que la colección cambie. En el caso de la búsqueda en repositorios de objetos de aprendizaje la colección de recursos es dinámica.

- **Expansión de términos basados en modelos de conocimiento independientes del corpus.**

  Por ejemplo el uso de tesauros u ontologías creadas a partir del conocimiento de los expertos o de los perfiles, historial y preferencias del usuario. La expansión con tesauros se realiza principalmente a través de relaciones léxicas de sinónimos, términos generales o términos específicos. Algunas propuestas al respecto se desarrollan en (Díaz-Galiano et al., 2009; Sangoi Pizzato & Strube de Lima, 2003; Sihvonen & Vakkari, 2004b). La expansión basada en ontologías por lo general se realiza a través de relaciones léxicas y con menos frecuencia, se utilizan relaciones ontológicas básicas como la relación padre/hijo (*is_a*). Algunas propuestas al respecto se desarrollan en (Ali & Khan, 2008; Calegari & Pasi, 2008; Lee et al., 2008; Ma et al., 2009; Jiuying Qin et al., 2009; Song et al., 2007; Tuominen et al., 2009).

Es preciso hacer notar que la mayoría de las técnicas propuestas utilizan ontologías de dominio o de tareas creadas para la propuesta, es decir no se utilizan ontologías concensuadas y validadas por expertos. Por otro lado, muy pocas de las ontologías utilizadas son formales, es decir se encuentran descritas a través de un lenguaje de representación formal. Más aún es común que en las investigaciones se plantee el uso de una ontología que según su propia definición sólo se trata de un tesauro.

3.4. Delimitación del ámbito del problema

Las principales características del escenario en el cual se enmarca esta investigación se resumen a continuación:

- La búsqueda en repositorios de OA se refiere a depósitos especializados para este tipo de recursos. Estos repositorios proveen funcionalidades pensadas para la gestión de este tipo de recursos, tanto de su contenido como de sus metadatos (IMS, 2003a; McGreal et al., 2008).
- El usuario de un repositorio de OA es distinto a un usuario Web "genérico". Por ejemplo la necesidad de búsqueda de un usuario Web es de naturaleza general. Es común que este usuario posea menos conocimiento en el dominio de la consulta y en consecuencia, sus consultas posean mayor nivel de ambigüedad (French et al., 2001). Nuestra investigación se centra en el diseñador instruccional, profesor o tutor quien busca recursos para integrarlos en el diseño de sus cursos. Hacer esta distinción permite destacar que este usuario posee un alto nivel de conocimientos en el dominio de la consulta, así como también conoce las restricciones respecto a la audiencia y las condiciones de uso de los recursos para alcanzar los aprendizajes esperados en el alumno.
- El diseñador instruccional consulta un repositorio para buscar objetos de aprendizaje. Un OA incluye el contenido propiamente tal y un conjunto de metadatos que especifican sus características. Si bien existen diferencias respecto al grado de compleción de los metadatos almacenados en distintos repositorios, por lo menos se mantiene un mínimo de información que facilita su localización (M.-Á. Sicilia et al., 2005). En el caso de la recuperación de información en la Web, el usuario puede requerir información de cualquier tipo, formato e incluso calidad, por ejemplo páginas, documentos, imágenes, videos, publicidad, etc.

En la Figura 22 se representan las características mencionadas anteriormente.

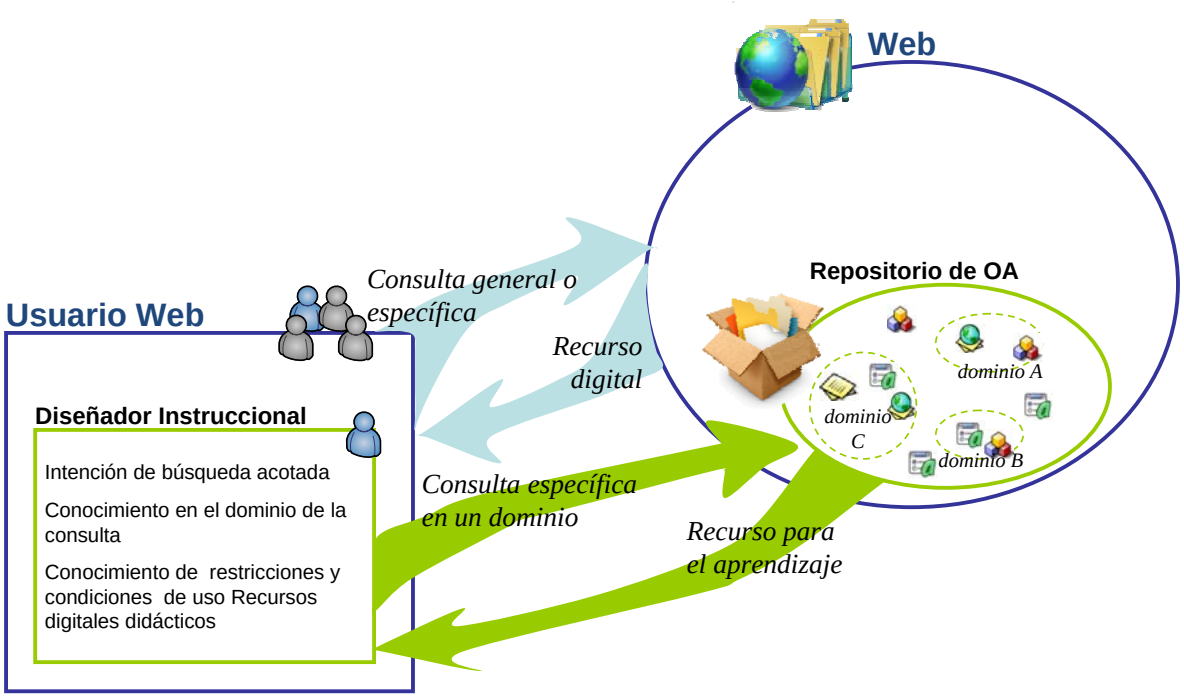


Figura 22: Recuperación de información Web versus búsqueda en repositorios de OA.

En síntesis, dadas las características de propuestas analizadas anteriormente, en la Tabla 10 destacamos las dimensiones en las cuales es posible hacer la diferencia con nuestra investigación.

Tabla 10: Síntesis de los aspectos diferenciadores de la invetigación.

Aspectos de diferenciación	Propósito de esta investigación
Tipo de modelo de representación de conocimiento utilizado	Ontología
Características del modelo de representación de conocimiento utilizado	Formal consensuada y validada por expertos
Tipo de relación utilizado para la expansión	Relaciones ontológicas
Usuario de la expansión	Diseñador instruccional
Naturaleza de la necesidad del usuario	Diseño o composición de recursos de aprendizaje
Tipo de recurso buscado	Objeto de aprendizaje
Sistema de RI	Repositorios de OA.

La expansión de consultas basadas en ontologías aplicada en la búsqueda en repositorios de objetos de aprendizaje está dirigida al *uso de ontologías formales de dominio, disponibles y validadas por expertos*. Específicamente, se propone *el uso de las relaciones ontológicas básicas* (por ejemplo, relaciones padres, hijos, hermanos, componentes y contenedores obtenidos de los tipos de relaciones *is_a, part_of* o *is_part*) y *relaciones ontológicas específicas* definidas para el dominio. Nuestros objetivos están dirigidos a apoyar al *diseñador instruccional* cuando éste busca *recursos de aprendizaje* en *repositorios especializados*. Por lo

tanto, se asume que la naturaleza de la necesidad de información del profesor es diferente a la naturaleza casual o general de las necesidades de información de un usuario en la Web.

3.5. Hipótesis de partida

La hipótesis de partida de esta tesis doctoral es la siguiente:

> *El uso de ontologías para la expansión de las consultas del diseñador instruccional en repositorios de objetos de aprendizaje puede mejorar la novedad de los resultados obtenidos.*

El objeto de estudio de esta tesis sólo concierne a recursos digitales formalmente almacenados (es decir, con sus metadatos) en repositorios de objetos de aprendizaje. Tampoco se considera el uso de diccionarios o tesauros generales, sino el uso de ontologías formales y consensuadas, es decir que se encuentren validadas por expertos y especificadas a través de un lenguaje de representación formal.

Dado que los diseñadores instruccionales son expertos en su dominio, el uso de ontologías del dominio como mecanismo para expandir las búsquedas puede servir para mejorar la recuperación de recursos en cuanto al acceso a recursos relevantes que sin la expansión no serían recuperados. En esta tesis se estudia de manera sistemática y empírica la expansión de consultas mediante ontologías en repositorios de OA.

3.6. Objetivos

El objetivo principal de esta tesis consiste en:

> *Proponer una estrategia para la expansión de consultas basada en ontologías de dominio que permita al diseñador instruccional obtener resultados relevantes desde los repositorios de objetos de aprendizaje.*

Se entiende que los objetos de aprendizaje relevantes obtenidos mediante esta estrategia no serían obtenidos sin la expansión. Para alcanzar este objetivo general, se plantean los siguientes sub-objetivos:

(O1) Revisar las técnicas de expansión de consultas, específicamente aquellas basadas en ontologías, y dentro de estas, las utilizadas o propuestas para la búsqueda en repositorios de objetos de aprendizaje.

(O2) Establecer los tipos de relaciones de la ontología y los tipos de expansión aplicables en el ámbito de la recuperación de objetos de aprendizaje en repositorios.

(O3) Especificar un procedimiento para resolver el problema de la correspondencia entre el concepto buscado y los conceptos modelados en la ontología en el

contexto de la búsqueda de objetos de aprendizaje.

(O4) Especificar una solución a los problemas o limitaciones detectadas para la implementación de la expansión de consultas basadas en ontologías en los repositorios de objetos de aprendizaje.

(O5) Plantear, diseñar e implementar la estrategia de expansión de consultas basadas en ontologías en el ámbito de la recuperación de objetos de aprendizaje en repositorios.

(O6) Comprobar la efectividad de la propuesta, su marco de aplicabilidad, sus limitaciones y sus posibilidades de extensión hacia otros dominios.

3.7. Método general

Para alcanzar los objetivos planteados, nuestra investigación será abordada a través de las siguientes fases:

1. Investigación y Análisis. Esta fase está dedicada a obtener la información que permita establecer el marco teórico de las distintas áreas de conocimiento implicadas en esta investigación. La recopilación y análisis será afrontada con un enfoque top-down, es decir partiendo desde las tres grandes áreas involucradas; la recuperación de información, los sistemas e-learning y la representación de conocimiento, hasta converger en la expansión de consultas basadas en ontologías sobre repositorios de objetos de aprendizaje.
2. Descripción del contexto del problema y estructuración de la investigación. Basado en los antecedentes del estado del arte, en esta fase se detallan los problemas detectados en la recuperación de OA en repositorios y se especifican las diferencias entre nuestra investigación y las propuestas previas para la expansión de consultas basadas en ontologías. Por último, se definen los aspectos que delimitan el contexto del problema y caracterizan nuestra investigación.
3. Definición y diseño de la propuesta. En esta fase se contempla la especificación de los criterios y/o restricciones para la expansión de consultas basada en ontologías. Se define la forma como serán abordados los problemas detectados en la fase anterior, y por último, se formula la estrategia de expansión de consultas basada en ontología de dominio aplicada a la búsqueda en repositorios de OA.
4. Evaluación. En esta fase, para evaluar la efectividad de la propuesta se diseña un experimento a partir del cual se obtendrá la relevancia resultados obtenidos con y sin la expansión. Se identifican las métricas de evaluación del desempeño en la recuperación de OA y los análisis estadísticos que respaldan la discusión de los resultados obtenidos.
5. Análisis de los resultados y formulación de conclusiones. En esta fase se analizan y discuten los resultados obtenidos en función de su marco de aplicabilidad, sus

limitaciones y sus posibilidades de extensión hacia otros dominios. Por último, se elaboran las conclusiones y se analiza el estado de consecución de los objetivos planteados. A partir de los hallazgos obtenidos de esta fase es posible establecer líneas de trabajo futuro.

Capítulo 4. Descripción de la solución

A continuación se detalla la propuesta de expansión de consultas basada en ontologías formales de dominio, aplicada en el contexto de la búsqueda de recursos de aprendizaje en repositorios especializados. Nuestra propuesta se plantea como solución al problema enunciado en la sección anterior.

4.1. Planteamiento de la solución propuesta

En general, la expansión de consultas basadas en ontología queda definida por el tipo de ontología, los tipos de relaciones utilizadas para la expansión y la distancia semántica de los nuevos términos respecto a la consulta inicial del usuario. A continuación se describen y justifican cada unos de estos aspectos.

4.1.1. Tipos de Relaciones

Los términos expandidos son extraídos desde las relaciones semánticas que el concepto buscado posee con otros conceptos modelados en la ontología. Algunas de estas relaciones son:

- Relaciones ontológicas básicas, entre las cuales se encuentran:
    - Relación *is_a*. Es una relación jerárquica y de herencia entre dos conceptos. En ella se indica que el concepto menor es un sub-tipo, una sub-clase del concepto padre del cual hereda sus propiedades. En el ejemplo presentado en la Figura 23 los conceptos *día-de-la-semana* y *día-de-visita-turística* son un tipo de la clase *temporal-región*.
    - Relación *part-of*. Es una relación jerárquica entre dos conceptos. En ella se indica que un concepto se compone de otros conceptos relacionados. En el ejemplo presentado en la Figura 23, una instancia de la clase *temporal-region* puede ser parte de otra instancia de la misma clase, es decir, el día *domingo* es

una instancia de la clase *días-festivos* y además forma parte de la clase *días-de-visita-turística*.

- Relaciones léxicas. Este tipo de relaciones son comunes en los diccionarios o tesauros, aunque también son incluidas en las ontologías para enriquecer la base de conocimiento. Algunos ejemplos de este tipo de relaciones son sinónimos, acrónimos, antónimos, términos relacionados y términos generales. En nuestra propuesta sólo se considera la relación léxica de sinonimia.

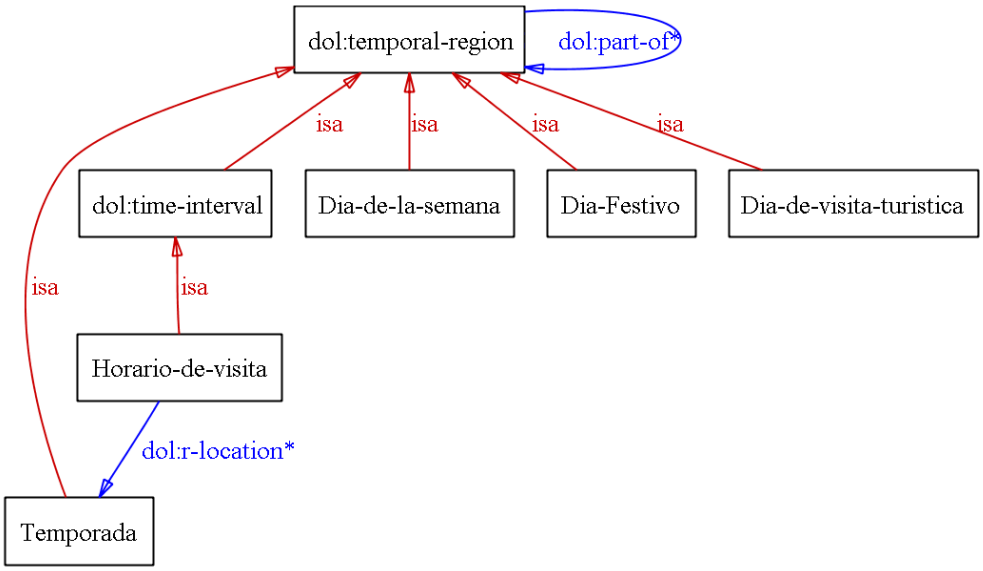


Figura 23: Ejemplo tipo de relaciones modeladas en la ontología de Turismo (ProtegeWiki, 2010).

El uso de estos dos tipos de relaciones en la expansión de consultas se basa en que son comunes en cualquier modelo de dominio. Aún cuando alguna de estas relaciones se encuentre etiquetada con un nombre diferente, el significado asociado a cada relación permite hacer su correspondencia con las relaciones presentes en cualquier modelo de dominio.

Además en la expansión se incluyen relaciones ontológicas específicas del dominio. Estas son tipos de relaciones válidas, sólo o especialmente, en un dominio. Por ejemplo en un dominio de turismo se modela la relación "*r-location*" que permite representar una "*zona abstracta*" dentro de otra zona. En el ejemplo presentado en la Figura 23 es posible modelar que el horario de visita a un sitio de interés se localiza dentro de una parte del día, tal como decir en la temporada de **verano** el horario de visita al **museo "el Prado"** es de **10 a 21** hrs.

4.1.2. Distancia semántica

La distancia semántica permite determinar el grado de cercanía entre los significados de dos conceptos (Budanitsky & Hirst, 2001; Jike & Yuhui, 2008). La medida de la distancia semántica es inversa al grado de relación semántica entre dos conceptos, es

decir a mayor distancia semántica entre el concepto `A` y `B`, menor es la cercanía o relación semántica entre ellos.

En una ontología la distancia semántica puede ser estimada en función de la longitud de la ruta (*path*) entre los dos conceptos modelados, aunque en la propuesta por (Jike & Yuhui, 2008) la ruta entre los conceptos es ponderada por la especificidad de los conceptos (distancia hacia la raíz). En otros casos la distancia es ponderada por los tipos de relaciones que conforman la ruta entre dos conceptos.

Como se observa en la Figura 24 la relación semántica entre los conceptos de la clase `jugos (juice)` y los conceptos de la clase `fruta(fuit)` es más lejana que la relación entre los conceptos de la clase `comestible (edible thing)` y los conceptos de la clase `fruta (fruit)`. En términos simples, es decir sin considerar el tipo de las relaciones involucradas o la especificidad de los conceptos, existe una mayor distancia semántica entre los conceptos `juice` y `fuit` que entre los conceptos `edible thing` y `fuit`:

`juice` → `potable liquid` → `consumable thing` → **`fruit`**

`edible thing` → `consumable thing` → **`fruit`**

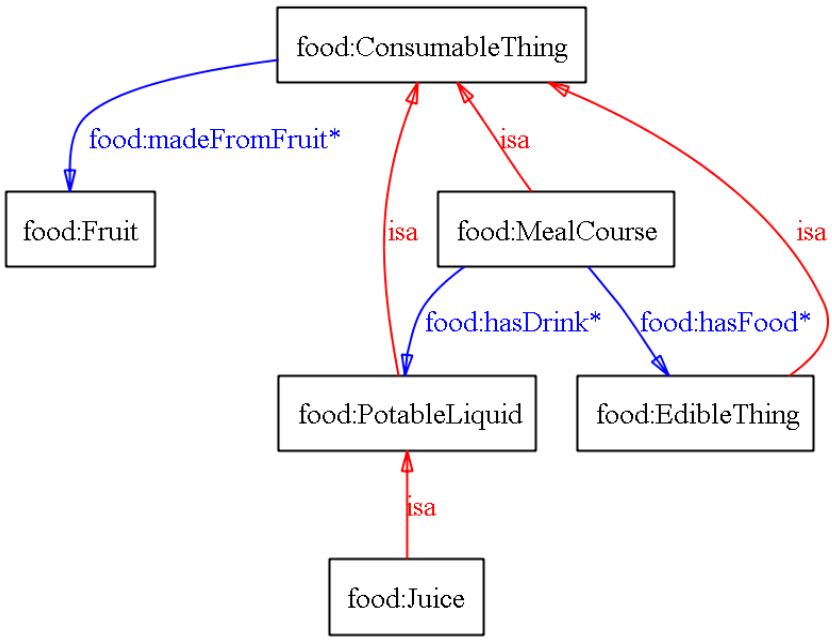


Figura 24: Ejemplo de distancia semántica utilizando parte de la ontología Food (ProtegeWiki, 2010)

Los nuevos conceptos que serán agregados a la consulta están restringidos a aquellos que poseen mayor grado de relación semántica, es decir que los conceptos seleccionados de la ontología para expandir la consulta están directamente relacionados con el concepto consultado. Esto implica que la expansión se ha restringido a los conceptos con una distancia semántica de 1 respecto al concepto consultado, independientemente del tipo de relación que exista entre ellos.

Dentro del contexto de esta investigación, es de destacar que el usuario es un diseñador instruccional que posee conocimientos en el dominio de la consulta, por lo tanto la ambigüedad de ésta es menor a la que se puede encontrar en la consulta de un usuario novel. Con la restricción de la distancia semántica igual a uno se intenta que los nuevos conceptos mejoren la especificidad a la consulta original y no agreguen ambigüedad innecesaria. En nuestro caso, la expansión con conceptos que se encuentran a mayor distancia semántica de la consulta inicial puede ampliar excesivamente el ámbito de la consulta definida por el usuario.

Téngase en cuenta que en la distancia semántica no se pondera el tipo de relación dado que es de nuestro interés analizar los resultados de las expansiones derivadas de cada uno de los tipos de relación existentes en la ontología.

4.1.3. Tipos de Expansiones

A partir de los tipos de relaciones descritos en la sección anterior, se definen los siguientes tipos de expansiones aplicables a las consultas del usuario:

- *is_a*: **Padres** (abreviado **isa_exa**). Se expande a través de los padres del concepto buscado. En este tipo de expansión se utiliza la relación ontológica *is_a*.
- *is_a*: **Hermanos** (abreviado **isa_her**). Se expande a través de los conceptos que comparten el mismo padre con el concepto buscado. Es decir la expansión se realiza con los conceptos hermanos. Si un concepto posee más de un padre se expanden todos los hijos de cada uno. En este tipo de expansión se utiliza la relación ontológica *is_a*.
- *is_a*: **Hijos** (abreviado **isa_hij**). Se expande a través de los conceptos que son hijos, es decir con todos los sub-tipos del concepto buscado. En este tipo de expansión se utiliza la relación ontológica *is_a*.
- *part_of*: **El todo** (abreviado **par_exa**). Se expande a través de los conceptos que contienen al concepto buscado, es decir se expande con él (o los) conceptos del cual el concepto buscado forma parte. En este tipo de expansión se utiliza la relación ontológica *part_of*.
- *part_of*: **Las partes** (abreviado **par_otr**). Se expande a través de los conceptos que están contenidos dentro del mismo concepto que contiene el concepto buscado. En este tipo de expansión se utiliza la relación ontológica *part_of*.
- **Sinónimo** (abreviado **syn_exa**). Se expande a través de los sinónimos del concepto buscado. Dependiendo de la ontología estas relaciones pueden incluir otros tipos de sinónimos, por ejemplo sinónimo exacto, relacionado, cercano, entre otros. En este tipo de expansión se utiliza la relación léxica de sinonimia.

Los tipos de expansiones definidos amplían o especifican la temática de la consulta. Por ejemplo las expansiones *is_a*: padres, *is_a*: hijos, *sinónimo*: exacto y *part_of*: el todo extraen los nuevos términos moviéndose verticalmente dentro de la temática de la consulta, lo

cual puede aportar mayor especificidad. Por otro lado, las expansiones *is_a*: hermanos y *part_of*: las partes extraen los nuevos términos moviéndose horizontalmente dentro de la temática de la consulta, es decir que amplían la temática aunque los nuevos términos siguen estando directamente relacionados con la consulta inicial (con una distancia semántica de uno).

Para ejemplificar cada uno de los tipos de expansión mencionados anteriormente, utilizamos la ontología GENE representada en la Figura 25 y Figura 26 (Diehl, Lee, Scheuermann, & Blake, 2007; Meng et al., 2009).

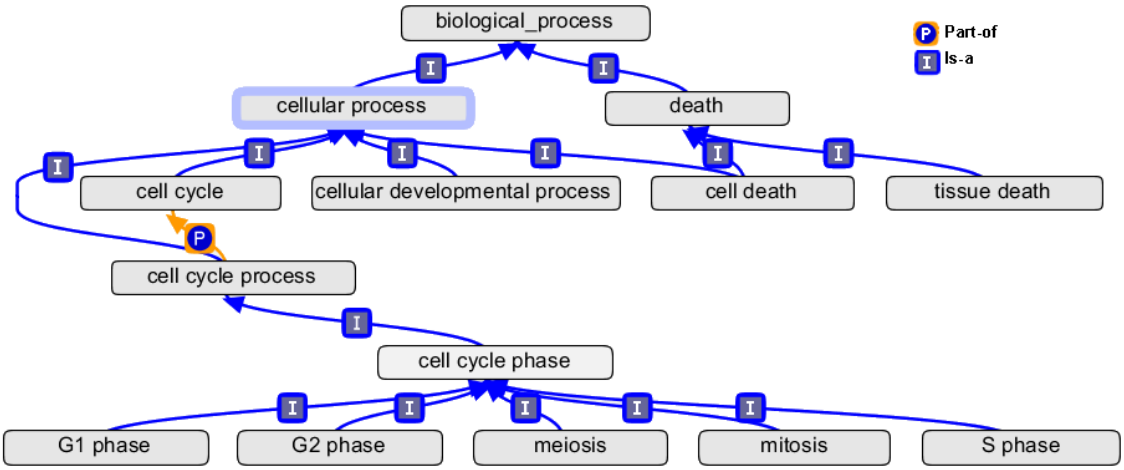


Figura 25: Parte de la ontología Gene donde se representan las relaciones del tipo *is_a* y *part-of* entre los conceptos representados. Imágenes capturadas mediante OBO-Edit software para la edición de ontologías. http://oboedit.org/.

- Expansión *is_a*: padres (*isa_exa*). Si el concepto buscado, es decir la consulta inicial del usuario es "*cell cycle*" la expansión *isa_exa* incluirá los padres de este concepto, en este caso "*cellular process*".

  Si la consulta inicial es "*cell death*" la expansión *isa_exa* incluirá los conceptos "*death*" y "*cellular process*", ya que dicho concepto posee 2 padres.

- Expansion *is_a*: hermanos (*isa_her*). Si el concepto buscado es "*cell cycle*" la expansión *isa_her* incluirá los conceptos "*cell death*", "*cellular developmental process*" y "*cell cycle process*". Estos tres conceptos son hijos del concepto "*cellular process*", que en este caso corresponde padre del concepto buscado.

- Expansión *is_a*: hijos (*isa_hij*). Si el concepto buscado es "*cell cycle phase*" la expansión *isa_hij* incluirá los conceptos "*G1 phase*", "*G2 phase*", "*meiosis*", "*mitosis*" y "*S phase*".

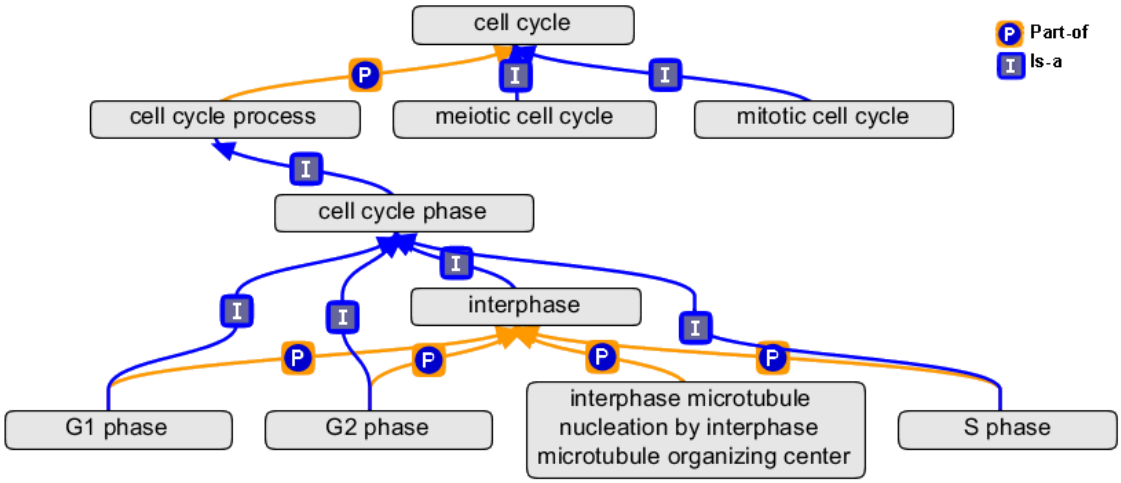


Figura 26: Parte de la ontología Gene donde se representan las relaciones del tipo *is_a* y *part_of* entre los conceptos representados.

- Expansión *part_of*: el todo (*par_exa*): Si el concepto buscado, es decir la consulta inicial del usuario es "*cell cycle process*" la expansión *par_exa* incluirá el concepto "*cell cycle*", ya que este concepto esta compuesto por el concepto buscado.
- Expansión *part_of*: las partes (*par_otr*). Si el concepto buscado es "*S phase*" la expansión *par_otr* incluirá los conceptos "*G1 phase*", "*interphase microtubule nucleation by interphase microtubule organizing center*" y "*G2 phase*", ya que estos conceptos también forman parte del concepto "*interphase*" que contiene al concepto buscado "*S phase*".

4.1.4. Solución al problema de la correspondencia en la ontología

La correspondencia entre los conceptos modelados en la ontología y los términos buscados en la consulta del usuario es uno de los mayores problemas cuando se utiliza una ontología para realizar la expansión. Para enfrentar esta situación algunos estudios proponen que la consulta sea formulada navegando en la ontología o utilizándola como un vocabulario controlado.

Debido a que en nuestra propuesta no se contempla la estrategia interactiva de formulación y refinamiento de las consultas, sino más bien una estrategia automática, en el caso que no exista una correspondencia (*matching*) exacta entre los conceptos de la ontología y la consulta, se utilizará el concepto que *más se acerca* al concepto buscado, a través de los siguientes pasos:

1. Extraer todas las *stopwords* de la consulta inicial.
2. Pre-seleccionar en la ontología los conceptos (sin *stopwords*) que contienen la mayor cantidad de las palabras del concepto buscado.
3. De la lista de conceptos preseleccionados se busca el concepto que incluye la mayor cantidad de palabras de la consulta y que además incluye la menor cantidad de palabras nuevas.

Por ejemplo, un profesor busca OA en el tema de genética con la consulta "*cell cycle regulation*", dado que dicho concepto no existe exactamente en la ontología Gene se busca el concepto más cercano, como sigue:

1. La consulta "*cell cycle regulation*" no contiene stopword, por lo tanto las palabras contenidas en el concepto buscado son "*cell*", "*cycle*" y "*regulation*".
2. En la ontología se pre-seleccionan los conceptos que contienen todas o parte de las palabras "*cell*", "*cycle*" y "*regulation*". Los conceptos pre-seleccionados son "*cell cycle checkpoint*", "*cellular process regulating host cell cycle in response to viruses*", "*modulation by symbiont of host cell cycle*", "*regulation of cell clycle*", "*negative regulation of cell cycle*", entre otros.
3. El concepto más cercano es "*regulation of cell clycle*", ya que es el concepto que contiene todas las palabras de la consulta y que agrega la menor cantidad de nuevos términos (en este caso la palabra *of* no es relevante puesto que es un *stopword*).

4.1.5. Solución al problema de la calidad del modelo de conocimiento

Nuestra propuesta plantea que la base de conocimiento utilizada para la expansión de consultas debe ser una ontología formal de dominio creada y validada con el consenso de los expertos.

Como fue mencionado anteriormente, el éxito de la expansión entre otras cosas depende de la calidad del modelo de conocimiento, es decir depende de la completitud, corrección, consistencia, coherencia y validez del conocimiento modelado. Asumimos que si la ontología utilizada es un modelo reconocido y respaldado por entidades de prestigio en el dominio, la calidad del conocimiento modelado es mayor.

4.1.6. Solución al problema en la generación de la cadena de búsqueda expandida

Tradicionalmente los términos expandidos son agregados a la consulta inicial a través del conector lógico OR (Billerbeck & Zobel, 2004). Por ejemplo el tipo de expansión *is_a*: hijos para la consulta inicial "*DNA replication*" obtiene 6 nuevos términos (ver la Tabla 11).

Tabla 11: Ejemplo de expansión, relación *is_a*: hijos.

Consulta original T3 "*DNA replication*", tipo de expansión *is_a*: hijos	
Términos Expandidos	1. DNA amplification
	2. DNA synthesis during DNA repair
	3. DNA-dependent DNA replication
	4. RNA-dependent DNA replication
	5. premeiotic DNA synthesis
	6. replication of extrachromosomal circular DNA

Para esta expansión, la cadena de búsqueda expandida quedaría expresada como:

```
 "DNA replication" OR "DNA amplification" OR "DNA synthesis
during DNA repair" OR "DNA-dependent DNA replication" OR
"RNA-dependent DNA replication" OR "premeiotic DNA synthesis"
OR "replication of extrachromosomal circular DNA"
```

Si bien, a través de su portal, algunos repositorios permiten la expresión de consultas utilizando conectores lógicos, esta facilidad no esta disponible en los servicios externos de consulta como los servicios RESTful o SQI. En estos casos no es posible tratar una cadena de búsqueda que contenga `n` conceptos y donde cada uno se componga de `m` palabras.

Considerando las restricciones mencionadas, si la expansión agrega `n` nuevos conceptos la cadena de búsqueda se divide en `n+1` (sub)consultas, es decir se ejecutarán las consultas por separado para el concepto inicial y los conceptos expandidos. Según el ejemplo anterior, la cadena de búsqueda se dividirá en siete consultas, una para el concepto original y seis para las expansiones, tal como sigue:

```
Q1:"DNA replication"
Q2:"DNA amplification"
Q3:"DNA synthesis during DNA repair"
Q4:"DNA-dependent DNA replication"
Q5:"RNA-dependent DNA replication"
Q6:"premeiotic DNA synthesis"
Q7:"replication of extrachromosomal circular DNA"
```

4.1.7. Solución al problema asociado al tratamiento de resultados duplicados

La medida tomada para resolver la problemática anterior, trae consigo una nueva problemática a resolver. Ésta tiene relación con el tratamiento de los resultados repetidos.

Si se aplican consultas por separado para cada uno de los conceptos expandidos, también se obtendrán listas de resultados independientes y en consecuencia, es muy probable que existan repeticiones entre los resultados recuperados para los conceptos expandidos de un mismo tipo de expansión y de una misma consulta.

Por cada consulta y tipo de expansión se debe obtener una única lista de resultados independiente de la cantidad de términos expandidos agregados a la consulta inicial. En el peor caso, cuando todos los conceptos consultados recuperan resultados en el repositorio, existen tantas listas de resultados como términos expandidos más uno (concepto inicial). En la Tabla 12 se presentan las listas de los resultados obtenidos para 2 de los 6 nuevos conceptos del tipo de expansión *is_a*: hijos para la consulta T3 "*DNA replication*". Téngase en cuenta que 4 de los conceptos expandidos para esta consulta no obtuvieron resultados en el repositorio.

Tabla 12: Lista de OA recuperados del repositorio (MERLOT) para los conceptos expandidos de la consulta T3 y el tipo de expansión *is_a*: hijos.

	Lista de los 10 primeros resultados para el concepto expandido: *DNA-dependent DNA replication*		Lista de los 10 primeros resultados para el concepto expandido: *RNA-dependent DNA replication*
0	DNA from the Beginning	**0**	**Transcription (DNA to RNA)**
1	Becoming whales - A lesson on whale evolution	**1**	**RNA Viruses**
2	DNA Structure 1.0	2	RNA Processing (post-transcriptional modifications)
3	**Transcription (DNA to RNA)**	3	RNA as the First Genetic Molecule
4	Nuclear Contents: DNA and Proteins	4	Biology Tutorials for Metabolism and Genetics
5	DNA in Bacteria and Viruses	5	Biology Flash Animations
6	DNA Replication (Semiconservative Replication)	6	Biology Workbench
7	The Genetic Code	7	The New Genetics
8	**RNA Viruses**	8	TranslationLab
9	DNA Double Helix (Watson and Crick)	9	World Index of Molecular Visualization Resources

Como se puede observar en las filas destacadas en color azul en la tabla, existen resultados repetidos y cada uno aparece en distintas posiciones del ranking. Para generar la lista final de los resultados del tipo de expansión *is_a*: hijos de la consulta T3 "*DNA replication*" se debe obtener un único ranking.

La solución planteada se basa en el supuesto de que "*la posición de un OA mejora si éste responde a la consulta de más de un término expandido*". En el ejemplo, el OA "*Transcription (DNA to RNA)*" es recuperado en 2 de las 6 consultas realizadas, en las posiciones de ranking 3 y 0, por lo tanto el ranking final de este OA debe ser mejor al de otros objetos recuperados sólo una vez.

Para calcular la posición final del ranking del j-ésimo OA proponemos la siguiente ecuación:

$$PLO_j = \begin{cases} x > 1; & \dfrac{Me\left(p_i LO_j\right)}{x} \\ x = 1 & p_i LO_j \end{cases}$$

Ecuación 9

Donde:
`j =1..t`, tal que `t` es el total de OA no repetidos.
`i=1..x`, tal que `x` es el numero de veces que el j-ésimo OA esta repetido en la lista.
$Me(p_i LO_j)$ es la mediana de las posiciones en las que el j-ésimo OA ha sido recuperado.

Esta ecuación toma como base la mediana de las posiciones de ranking que ha obtenido cada OA en cada repetición. Dado que la mediana es la posición de ranking que deja el mismo número de datos antes y después que él, una vez que estos han sido ordenados, consideramos que esta posición de ranking es más representativa del conjunto. Es de recodar que el conjunto de `x` posiciones de ranking representa el hecho que el mismo

OA fue recuperado como relevante cuando se aplicó la consulta de `x` conceptos expandidos.

También es posible utilizar otras medidas en vez de la mediana, tales como la media, la mínima o la máxima posición de ranking. El uso de la posición mínima o máxima nos lleva a un escenario optimista o pesimista, respectivamente. Por su parte, los estadísticos de centralidad como la media y la mediana nos señalan un escenario más probable. Aunque existen casos en los que el valor de la mediana y la media serán iguales también existen situaciones en los que la media es menos apropiada para representar nuestro conjunto de datos.

Por ejemplo si un OA ha sido recuperado al consultar 3 conceptos (`x=3`) en las posiciones de ranking `{9,1,1}`, se observa que en más de la mitad de los casos la posición es 1, es decir que el OA es evaluado como el más relevante en las listas de resultados recuperados. Para este ejemplo, la media de las posiciones de ranking es `4` y la mediana es `1`, donde este último consideramos que es un valor más probable y representativo del conjunto.

Una vez re-calculada la posición de ranking de las instancias repetidas, la lista final se obtiene enumerando a partir de uno (`1`) los resultados ordenados según el valor de PLO_j. Para el ejemplo desarrollado anteriormente, los valores obtenidos de `PLO` se presentan en la Tabla 13, los resultados destacados en color azul corresponden a los OA repetidos y a los cuales se les re-calculó su posición en el ranking.

Tabla 13: Resultados de la re-ponderación y eliminación de resultados repetidos (objetos de aprendizaje).

Orden[b]	PLO[a]	Título del Objeto de aprendizaje
1	0	DNA from the Beginning
2	**0,75**	**Transcription (DNA to RNA)**
3	1	Becoming whales - A lesson on whale evolution
4	2	RNA Processing (post-transcriptional modifications)
5	2	DNA Structure 1.0
6	**2,25**	**RNA Viruses**
7	3	RNA as the First Genetic Molecule
8	4	Biology Tutorials for Metabolism and Genetics
9	4	Nuclear Contents: DNA and Proteins
10	5	Biology Flash Animations

a: Posición final del ranking del OA
b: Posición en la lista de resultados

En la Tabla 14 se muestra un ejemplo con la lista de OA resultantes de la consulta T3 "*DNA replication*" y sus expansiones del tipo *is_a*: hermanos e *is_a*: hijos. Los resultados destacados en color azul corresponden a los OA repetidos entre los resultados de distintos tipos de expansión.

En resumen, como se representa en la Figura 27 cada consulta obtendrá una lista individual de resultados por cada uno de los tipos de expansión aplicado (*lista individual por tipo de expansión*). A partir de estas listas se obtiene la lista de resultados que será presentada y evaluada por el usuario (*lista por tipo de expansión*).

Tabla 14: Listas individuales de los 10 primeros OA recuperados por cada tipo de expansión aplicado a la consulta T3.

	Lista OA Expansión *is_a* : hermanos
1	**Transcription (DNA to RNA)**
2	Significance of DNA Sequence
3	Repair of Genetic Mutations
4	Ojala que llueva cafe
5	**DNA Structure 1.0**
6	Kimball's Biology Pages
7	**RNA Processing (post-transcriptional modifications)**
8	Entrez
9	eStroke Animated Chinese Characters
10	Computer Number Base Conversion Program
…	

	Lista OA Expansión *is_a* : hijos
1	DNA from the Beginning
2	**Transcription (DNA to RNA)**
3	Becoming whales - A lesson on whale evolution
4	**RNA Processing (post-transcriptional modifications)**
5	**DNA Structure 1.0**
6	RNA Viruses
7	RNA as the First Genetic Molecule
8	Biology Tutorials for Metabolism and Genetics
9	Nuclear Contents: DNA and Proteins
10	Biology Flash Animations

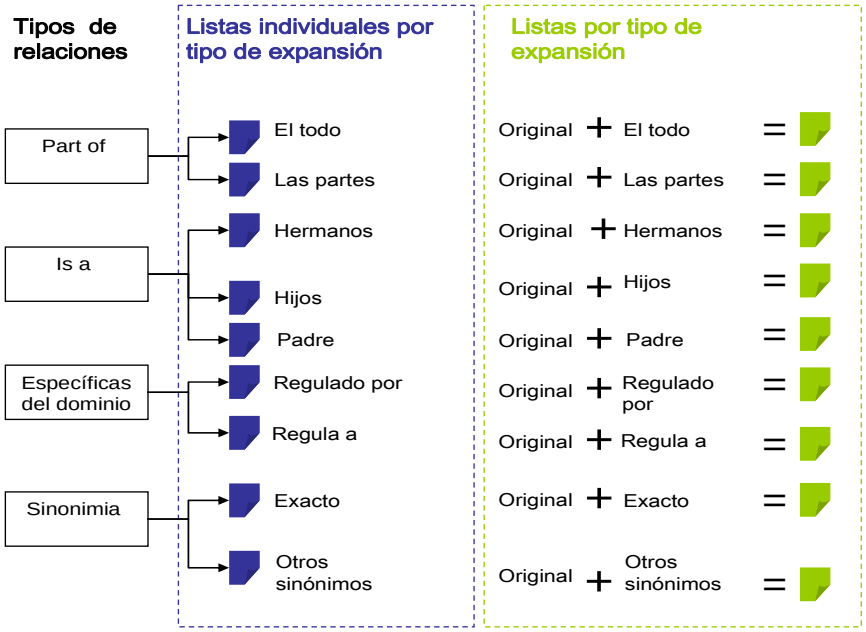


Figura 27: Esquema para la generación de las listas de resultados por tipo de expansión

Una vez obtenidas las listas de los resultados por tipo de expansión, es relevante analizar también los resultados de la expansión independiente de su tipo. Para esto se debe formar una lista única que reúna los resultados de todos los tipos de expansiones aplicados a la consulta. La *lista única de expansión* es obtenida de la intercalación de las *listas individuales por tipo de expansión* aplicadas a cada consulta. En la Figura 28 se representa la forma como se obtienen cada una de las listas mencionadas.

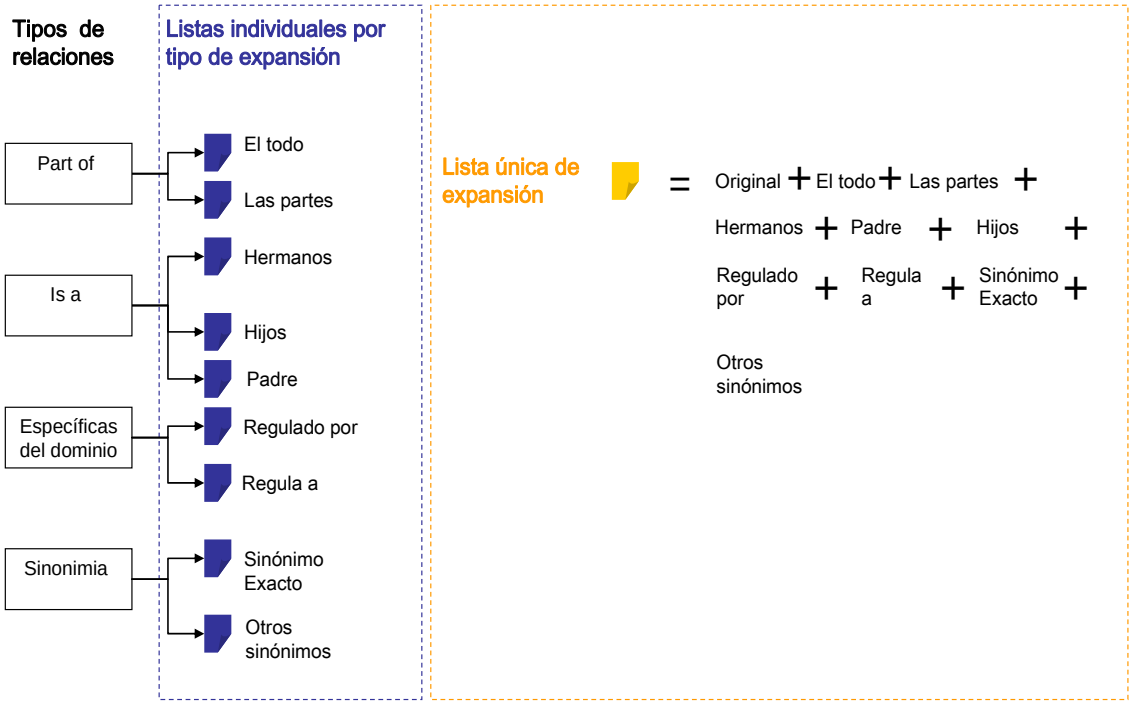


Figura 28: Esquema para la generación de las listas únicas de expansión.

Por ejemplo la consulta T16 obtuvo resultados en 3 tipos de expansión, *is_a*: hermano, *part_of*: el todo y *part_of*: las partes. Por cada uno de ellos existe una *lista individual por tipo de expansión*. A partir de la unión entre ellas y los resultados de la consulta original se obtienen las *listas por tipo de expansión*, para la consulta T16 son:

```
Listas por tipo de expansión de la consulta T16:

Original +isa her
Original +par exa
Original +par_otr
```

Finalmente la *lista única de expansión* se conforma de la unión de la lista de resultados de la consulta original y las *listas individuales por tipo de expansión* aplicados a la consulta. Para la consulta T16 las *listas únicas de expansión* son cualquiera de sus 6 permutaciones, `n!=3!=6`, donde `n` es la cantidad de tipos de expansión aplicados a la consulta:

```
Listas únicas de expansión de la consulta T16:

Original +isa her +par exa +par otr
Original +isa_her +par_otr +par_exa
Original +par_exa +par_otr +isa_her
```

```
Original +par exa +isa her +par otr
Original +par otr +par exa +isa her
Original +par_otr +isa_her +par_exa
```

En el caso de la consulta T3 se aplicaron sólo dos tipos de expansiones, i*s_a*: hermanos (isa_her) y *is_a*: hijos (isa_hij), por lo cual las *listas únicas de expansión* son

```
Listas únicas de expansión de la consulta T3:
Original +isa hij +isa her
Original +isa_her +isa_hij
```

Al realizar la unión de las listas es posible encontrar OA repetidos entre las listas, en este caso se asumirá un criterio optimista manteniendo el mejor ranking entre las repeticiones. Siguiendo el ejemplo de la consulta T3 presentado en la Tabla 14, los OA destacados son aquellos que fueron recuperados para más de un tipo de expansión, al ser incluidos en la lista final sólo se considera aquel OA con mejor posición de ranking.

En la Tabla 15 se presentan las dos opciones para la lista única de la expansión de la consulta T3, la diferencia entre ellas radica en la importancia que tiene cada tipo de expansión a la hora de realizar la unión de los resultados, es decir el orden de intercalación de éstos. La primera lista pondera mejor los resultados de la expansión *is_a*: hijos que los resultados de la expansión *is_a*: hermanos. Por el contrario, la segunda lista otorga mayor importancia a los resultados de la expansión *is_a*: hermanos que a los resultados de la expansión *is_a*: hijos.

Tabla 15: Listas de OA únicas. Ambas listas ponderan mejor un tipo de expansión distinto.

	Lista OA Consulta original (sin expansión)
1	DNA Replication (Semiconservative Replication)
2	Biology Tutorials for Metabolism and Genetics
3	DNA Replication Video
4	Lew-Port DNA Replication
5	Biology Flash Animations
6	Lew-Port's Biology Place--Animated Reviews
7	DNA Workshop
8	DNA Replication
9	WEHI-TV DNA Molecular Animation
10	Animation of Replication

	Lista OA Consulta original**+*is_a*: hermanos+*is_a*: hijos**
1	DNA Replication (Semiconservative Replication)
2	Transcription (DNA to RNA)
3	DNA from the Beginning
4	Biology Tutorials for Metabolism and Genetics
5	Significance of DNA Sequence
6	DNA Replication Video
7	Repair of Genetic Mutations
8	Becoming whales - A lesson on whale evolution
9	Lew-Port DNA Replication
10	Ojala que llueva cafe

	Lista OA Consulta original**+*is_a*: hijos+*is_a* : hermanos**
1	DNA Replication (Semiconservative Replication)
2	DNA from the Beginning
3	Transcription (DNA to RNA)
4	Biology Tutorials for Metabolism and Genetics
5	Significance of DNA Sequence
6	DNA Replication Video
7	Becoming whales - A lesson on whale evolution
8	Repair of Genetic Mutations
9	Lew-Port DNA Replication
10	RNA Processing (post-transcriptional modifications)

4.1.8. Resumen del Planteamiento

La estrategia de expansión de consultas basada en ontologías aplicada a la búsqueda de OA en repositorios se basa en los siguientes criterios:

- Tipo de expansión automática. Es decir el usuario no interactúa con el modelo de conocimiento y tampoco interviene en la selección de los términos expandidos.
- Uso de ontologías formales de dominio, validadas o reconocidas por la comunidad disciplinaria y modeladas a través de un lenguaje de representación formal.
- Los nuevos términos son extraídos a partir de las relaciones ontológicas básicas y específicas de dominio, y las relaciones léxicas de sinonimia.
- Los conceptos consultados se relacionan con los conceptos expandidos con una distancia semántica de 1.
- La falta de correspondencia entre los conceptos consultados y los conceptos modelados en la ontología se resuelve buscando el concepto más cercano y que agregue menor ambigüedad a la consulta.
- Los nuevos términos no están restringidos en cuanto a cantidad y tampoco a su importancia dentro de la cadena de búsqueda.
- Los nuevos términos son agregados a la consulta inicial como una disyunción lógica (OR).
- Las repeticiones dentro de la lista de resultados por tipo de expansión son eliminadas y la nueva posición de ranking se estima considerando la mediana de las posiciones de los resultados repetidos (detectados en los OA recuperados por los conceptos expandidos para un mismo tipo de expansión y consulta).
- Las repeticiones dentro de la lista única de expansión son eliminadas y la nueva posición en el ranking se estima considerando la mejor de las posiciones de los resultados repetidos (detectados en los OA recuperados por los distintos tipos de expansión aplicados a una consulta).

4.2. Diseño de la solución

En la Figura 29 se presenta la arquitectura funcional que da soporte a la propuesta de expansión de consultas basada en ontologías dentro del contexto de la búsqueda de objetos de aprendizaje en repositorios. Esta arquitectura se divide en 4 grandes componentes: expansión de consultas basada en ontología de dominio, gestor de bases de conocimiento del dominio, búsqueda en repositorios y manejador de resultados.

La arquitectura propuesta responde a la solicitud de recursos hecha por el profesor a través de una interfaz web de usuario o puede responder a la solicitud de expansión hecha por un LMS, una federación de repositorios o un repositorio individual. En este último caso sólo intervienen dos componentes: de expansión de consultas basada en ontología de dominio y el gestor de bases de conocimiento del dominio, siendo el emisor de la solicitud quien se encarga de la búsqueda y el procesamiento de los resultados.

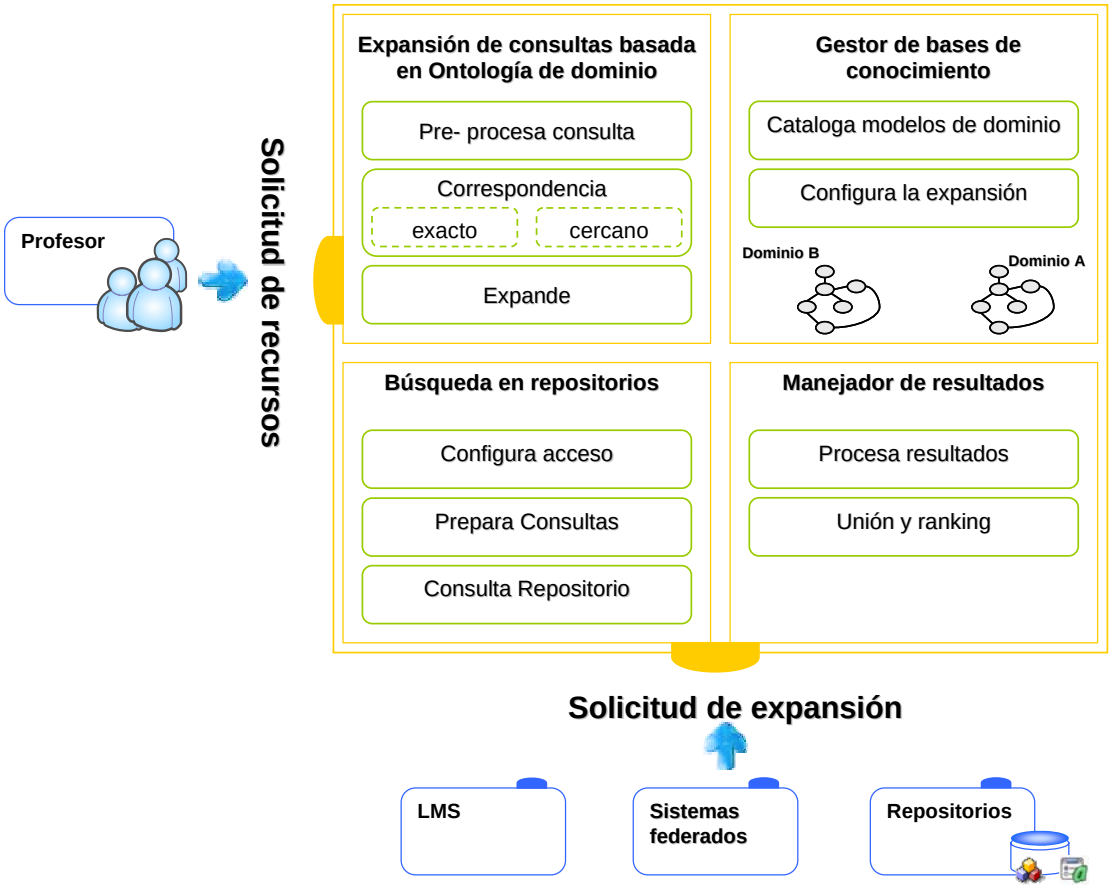


Figura 29: Arquitectura funcional que da soporte a la expansión de consultas basada en ontologías en el contexto de la búsqueda de OA en repositorios.

4.2.1. Arquitectura técnica

Para la implementación de los módulos *off-line* se utiliza JAVA como lenguaje de programación y framework de soporte. Además, se analizaron e integraron diferentes herramientas, *frameworks* y APIs entre las que se destacan:

- Lenguaje de consulta SPARQL para RDF. SPARQL puede ser usado para expresar consultas a través de diversas fuentes de datos.

Los resultados de las consultas SPARQL pueden ser tratados como *ResultSets* o como gráficos de RDF.

- Protégé. Es un editor de ontologías y modelos de conocimientos. Permite crear y mantener ontologías, por ejemplo definir clases, relaciones, instancias, propiedades y eventos de razonamiento (Stanford Center for Biomedical Informatics Research at the Stanford University School of Medicine, 2010).

- Jena. Es un marco de trabajo Java para desarrollar aplicaciones de Web Semántica. Suministra un ambiente de programación para RDF, RDFS, OWL, SPARQL y además incluye un motor de inferencia basado en reglas (Hewlett-Packard Development Company, 2008).

  Jena incluye:

  - La API de RDF.
  - Lectura y escritura RDF en RDF / XML, N3 y N – tripletas.
  - La API de OWL.
  - Almacenamiento persistente y en memoria.
  - Motor de consultas de SPARQL.

- ARQ Query Processor. ARQ es un motor de consultas para Jena que permite ejecutar consultas SPARQL. Dentro de los paquetes contenidos se destacan:

  - com.hp.hpl.jena.query
  - com.hp.hpl.jena.query.larq
  - com.hp.hpl.jena.update

- Lucene Apache. Es una librería escrita completamente en Java que implementa un motor de búsqueda de texto. Dentro de los paquetes implementados en Lucene se destacan:

  - org.apache.lucene.analysis.standard. Este paquete es un analizador gramático estándar que se utiliza para implementar los análisis básicos de las consultas y documentos a indexar.
  - org.apache.lucene.index. Este paquete permite la mantención y creación de índices.
  - org.apache.lucene.queryParser. Este paquete es un analizador sintáctico de consulta que se utiliza para la eliminación de stopword, de caracteres especiales y de espacios vacíos, entre otros.
  - org.apache.lucene.search. Este paquete permite la búsqueda en índices.
  - org.apache.lucene.util. Este paquete incluye herramientas adicionales, utilizadas, por ejemplo, para el manejo de los resultados.

- JDOM. Es un API para leer, crear y manipular documentos XML (JDOM TM Project, 2009).
- Xpath. *XML Path Language* es un lenguaje que permite recorrer y procesar un documento XML a través de expresiones. XPath permite buscar y seleccionar teniendo en cuenta la estructura jerárquica del XML (Clark & DeRose, 1999).

Para la implementación del módulo Web que funciona en modo *on-line* se utilizó Php sobre apache Tomcat 5.5.

4.2.2. Descripción funcional del prototipo de implementación

La solución planteada en la sección 4.1 se implementa en un prototipo que funciona en dos modos de operación, *off-line* y *on-line*. En la Figura 30 se representa la arquitectura técnica y funcional del prototipo de implementado en esta tesis. Los componentes achurados en la figura no fueron implementados en el prototipo.

El prototipo que implementa la solución se divide en los siguientes módulos:

- **Pre-procesamiento de las consultas**.

  Cada una de las consultas de prueba pasa primero por un proceso de limpieza donde se transforma a minúsculas, se eliminan símbolos y caracteres especiales.

  Este proceso se realiza *off-line* tomando secuencialmente cada una de las consultas de prueba que se encuentran en un archivo de texto plano, tal como se muestra en el extracto siguiente, cada consulta está separada por ";".

```
population genetic;
Mutation;
DNA replication;
gene-expression;
recombination;
"transposable element";
developmental genetic;
```

  La salida de este módulo es también un archivo de texto plano, donde las consultas pre-procesadas están separadas por un ";".

```
population genetic;
mutation;
dna replication;
gene expression;
recombination;
transposable element;
developmental genetic;
```

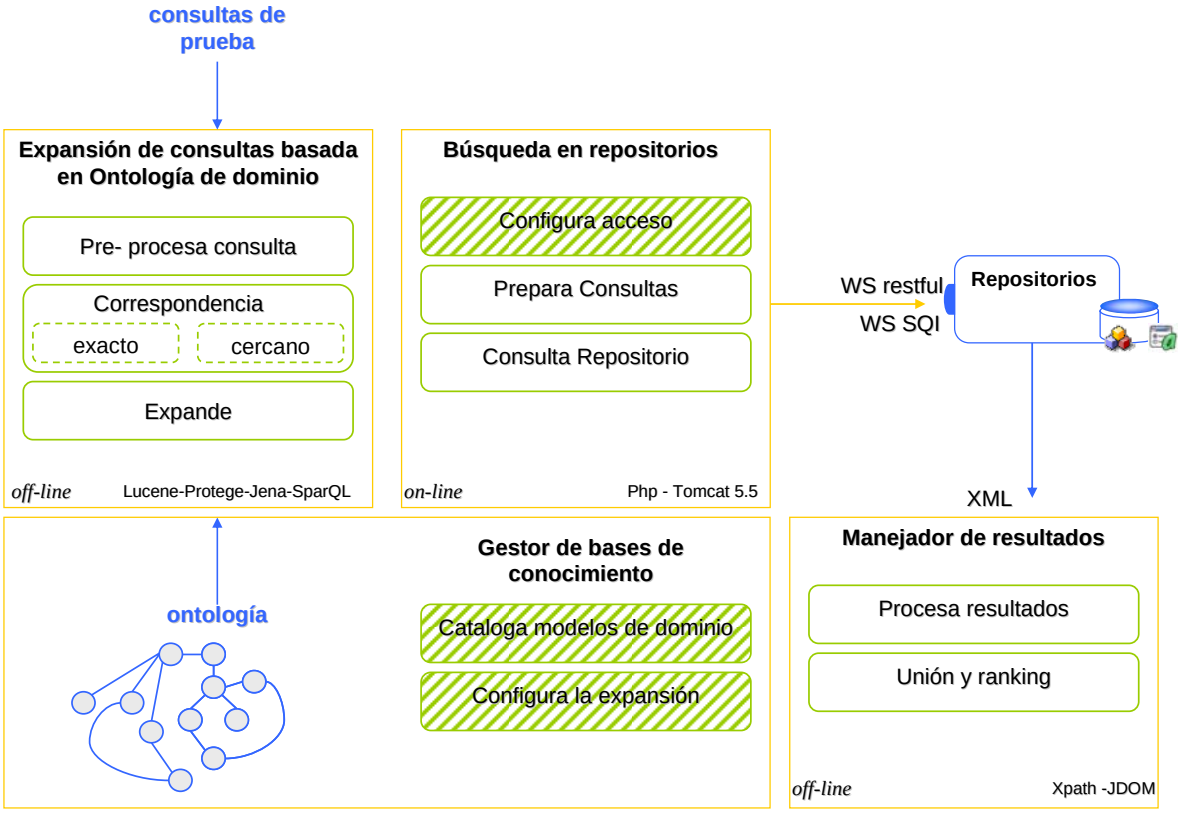


Figura 30: Arquitectura técnico-funcional del prototipo de implementación.

- **Correspondencia exacta y aproximada**.

  Una vez que la consulta inicial ha sido pre-procesada se realiza la búsqueda del concepto de la consulta en la ontología, es decir se busca la correspondencia exacta *(matching)* entre la consulta y los conceptos modelados en la ontología.

  Como fue explicado anteriormente, es posible que no exista una coincidencia exacta, en dicho caso se busca en la ontología el concepto más cercano a la consulta inicial. Se extraen de la ontología los conceptos que contienen al concepto buscado (sin *stopwords*) y luego se selecciona el concepto que agrega la menor cantidad de palabras nuevas a las contenidas en el concepto buscado. Es decir se selecciona el concepto que representa con menor ambigüedad a la consulta inicial. Para efectos de análisis, se guarda un registro con los conceptos que no existen exactamente en la ontología y tienen un concepto aproximado.

  La salida de este módulo es un archivo de texto plano con 4 campos, tal como se representa en el extracto siguiente están los campos: un identificador de la consulta (sólo de uso interno), la cadena de búsqueda, el identificador en la ontología y el concepto cercano en el caso que corresponda. Si sólo el último campo es NULL indica que la consulta tiene una correspondencia exacta y si los dos últimos campos son NULL indica que el concepto consultado no existe en la ontología y tampoco posee un concepto cercano. Si el campo "aprox" es distinto de NULL indica que el

concepto buscado no existe de forma exacta pero posee un concepto cercano a través del cual realizar las expansiones.

```
idQ; labelquery;              id_onto;    Aprox

T1;  population genetic;     NULL;       NULL
T3;  dna replication;        GO_0006260; NULL;
T5;  recombination;          GO_0006310; DNA recombination
```

- **Expansión**.

  Una vez encontrado el concepto buscado en la ontología, ya sea de forma exacta o aproximada, se ejecutan los distintos tipos de expansiones: *is_a*: padres, *is_a*: hijos, *is_a*: hermanos, *part_of*: el todo, *part_of*: las partes, *sinónimo*: exacto, etc.

  Por cada tipo de expansión se obtendrán desde cero a `n` nuevos términos. La cantidad máxima sólo queda determinada por la cantidad de relaciones existentes entre el concepto buscado y los demás conceptos modelados en la ontología. La cantidad de términos es cero si en la ontología el concepto buscado no posee el tipo de relación necesario para hacer la expansión.

  El resultado de este proceso es un archivo de texto plano donde se registran algunos campos importantes tales como el concepto buscado, el concepto aproximado, el identificador del concepto, el tipo de expansión aplicada y la lista de los términos expandidos.

```
idQ; labelquery; id_onto; Aprox;expan_type;exp-term...

T1;population genetic;NULL;NULL;no existe;NULL
T3;dna replication;GO_0006260;NULL;syn_all;DNA
biosynthesis;DNA biosynthetic process;DNA synthesis;
T5;recombination;GO_0006310;DNA recombination;a_par_exa;NULL
```

  En el extracto presentado antes tenemos que:

  - La consulta T1 "*population genetic*" no existe en la ontología y tampoco posee un concepto cercano.
  - Luego la consulta T3 "*dna replication*" existe en la ontología y su código es GO_0006260, los términos expandidos como sinónimos son "*DNA biosynthesis; DNA biosynthetic process; DNA sintesis*".
  - Por último, la consulta T5 "*recombination*" contiene un valor para el campo **Aprox** lo que indica que el concepto inicial no existe en la ontología y su concepto más cercano es "*DNA recombination*". En este caso no se obtuvieron nuevos términos para el tipo de expansión *part_of*: el todo (par_exa).

- **Consulta al repositorio**.

  Debido a las restricciones encontradas en los repositorios en cuanto al lenguaje de las consultas, las consultas expandidas son separadas en tantas subconsultas como

conceptos expandidos hayan sido agregados. La cadena de búsqueda de cada subconsulta contiene sólo un concepto, no obstante dado que un concepto puede estar formado de `m` palabras, la búsqueda es tratada como una "frase exacta".

Una vez preparadas las consultas, se establece la conexión con el repositorio y se envían a éste para su procesamiento.

- **Procesamiento de los resultados**.

  Independiente del sistema de recuperación utilizado por el repositorio para procesar la consulta y obtener el ranking de los objetos de aprendizaje que la satisfacen, el repositorio responde con un archivo XML donde se listan los metadatos de los OA recuperados.

- **Unión y Ranking de los resultados**.

  Dado que las consultas expandidas fueron separadas en subconsultas y que además por cada una de ellas existe una lista de resultados, es necesario unir las listas de OA, eliminar los OA duplicados, re-calcular la posición de los OA repetidos y generar la lista de resultados por cada tipo de expansión.

En el Anexo D se incluye la documentación de las clases de Java que implementan el módulo de expansión.

Capítulo 5.Evaluación

Una vez especificada e implementada la propuesta de solución, a continuación se detalla el experimento a través del cual es posible evaluar los resultados obtenidos de la estrategia de expansión especificada en el capitulo anterior y evaluar la hipótesis de nuestra investigación.

5.1. Contexto

Nuestro caso de estudio parte de la base que las consultas representan una necesidad de un diseñador de instrucción cuando prepara un curso de e-learning. Dado que no existe una colección estándar de prueba en este ámbito, para llevar a cabo la evaluación de nuestra propuesta es preciso definir el área de conocimientos, seleccionar el repositorio, definir el conjunto de consultas de prueba, y posteriormente, realizar la evaluación de la relevancia de los resultados.

El área de conocimientos en la cual evaluar nuestra propuesta requiere que exista una ontología de dominio formal y una colección de OA en el área. En nuestro caso de estudio, utilizamos la ontología GENE como la base del conocimiento de expertos para la expansión de consultas. Esta ontología ha sido validada por la comunidad científica en el área genética. Además, en el repositorio MERLOT existe una colección de OA en dicha área que nos permitirá aplicar las consultas de prueba.

El conjunto de consultas de prueba se extrae desde la lista de los contenidos tratados en los programas académicos de los cursos en el área de genética.

5.1.1. Ontología Gene

La ontología GENE (GO) es utilizada como la base de conocimiento experto a través del cual realizar la expansión de consultas. El proyecto GENE comienza en 1998, como colaboración entre las bases de datos de organismos como; *Flybase* (*Drosophila*), the *Saccharomyces Genome Database (SGD*) y *the Mouse Genome Database (MGD)*. Esta ontología es respaldada por el consorcio GO y la subvención *P41 de National Human Genome Research Institute* (NHGRI).

La Ontología Gene[15] esta compuesta a su vez de otras 3 ontologías (Diehl et al., 2007; Meng et al., 2009):

- *Cellular Component*. La ontología de componentes celulares puede ser usada para describir los productos genéticos de acuerdo a su localización o como parte de una proteína compleja. Esta ontología contiene 255.278 productos genéticos.
- *Molecular function*. La ontología de función molecular incluye los términos que describen las actividades enzimáticas y vinculantes. Esta ontología contiene 288.391 productos genéticos.
- *Biological Process*. La ontología de procesos biológicos contiene los términos que describen una serie de eventos que ocurren en una célula u organismo generados por una o mas funciones moleculares. Esta ontología contiene 270.683 productos genéticos.

Los aspectos más importantes que justifican la elección de esta ontología en la evaluación de nuestra propuesta son:

- Gene es una ontología de dominio validada por la comunidad científica en el área.
- Gene se encuentra disponible en formato OWL (RDF-XML) lo que permite el procesamiento automático de expansión de las consultas.
- Existe una colección de objetos de aprendizaje en este dominio que permite su uso en un repositorio.

5.1.2. Repositorio MERLOT

MERLOT fue desarrollado por el *Center for Distributed Learning* de la *California State University*. Este es un de los repositorios más maduro y reconocido por la comunidad e-learning, a la fecha[16] cuenta con 75.157 miembros y más de 21.396 materiales (Ochoa & Duval, 2009).

A diferencia de otros repositorios, MERLOT fue uno de los primeros en implementar un enfoque para la evaluación de la calidad de los OA sobre la base de un proceso de revisión por pares (Nesbit, Belfer, & Vargo, 2002) Hoy en día, el sitio Web de MERLOT da soporte más de veinte comunidades propias de la disciplina (por ejemplo, biología, historia, tecnología de la información, formación del profesorado, etc.). Cada comunidad disciplinaria tiene asociado un subconjunto de la colección completa de OA y cuenta con un consejo editorial que guía las políticas de revisión por pares y las prácticas en esa comunidad. De acuerdo con (Ochoa & Duval, 2009) MERLOT tiene la mayor base de contribuyentes voluntarios de recursos, alrededor de 1.446 (McMartin, 2004).

---

[15] http://www.geneontology.org/index.shtml?all, consultado en 26th Octubre 2009

[16] http://www.MERLOT.org/MERLOT/index.htm , consultado en 26th Octubre 2009

Para efectos de nuestra investigación y específicamente, para la evaluación de la propuesta, MERLOT es uno de los pocos repositorios que posee una colección específica más de doscientos recursos para el aprendizaje catalogados en el área genética dentro de la comunidad disciplinaria de biología. Desde el punto de vista técnico, MERLOT provee, además del portal Web, otros mecanismos de consulta off-line que facilitan la ejecución de nuestro experimento (Smith, Smith, & Melder, 2007). Por ejemplo el servicio Web RESTful o la especificación WSDL[17] (Christensen, Curbera, Meredith, & Weerawarana, 2001) para la implementación de un servicio Web cliente SQI (CEN, 2005). Dentro de estas dos opciones, el servicio RESTful provee mayores facilidades para restringir y especificar la consulta y el acceso a los OA recuperados, en la sección 2.2.4.1 se encuentran especificadas estas opciones en detalle.

5.2. Proceso

En la Figura 31 se representa el proceso seguido para contrastar la hipótesis inicial de la investigación.

El proceso comienza con la definición de las consultas que serán utilizadas en el caso de estudio. Como ha sido mencionado anteriormente, esta investigación se concentra en el diseñador instruccional como usuario, específicamente cuando éste enfrenta la tarea de diseñar o componer un nuevo OA o un curso e-learning. Por lo tanto, las consultas representan las necesidades de recursos en un determinado tema. De acuerdo a lo anterior, las consultas fueron definidas a partir de los programas de los cursos oficiales para las asignaturas impartidas por instituciones de educación (*syllabus*). Específicamente, las consultas fueron extraídas de los *syllabus* de cursos de genética impartidos por instituciones de educación superior, y publicados en la Web para el año 2009.

El siguiente paso en el proceso es realizar la expansión de las consultas de prueba. Debido a que las consultas no fueron extraídas desde la misma ontología, en esta etapa se presentan tres situaciones:

- El concepto consultado existe exactamente en la ontología,
- El concepto consultado no existe en la ontología, por lo tanto se busca el concepto más cercano, y
- Si no se encuentra un concepto cercano no es posible hacer expansión.

[17] WSDL es un formato XML para describir servicios web.

Una vez encontrado el concepto en la ontología (de forma exacta o aproximada) se realiza la expansión dependiendo de las relaciones que se encuentren modeladas en la ontología. Luego, se formulan las consultas y se interroga al repositorio. Los resultados para las consultas originales y sus expansiones son procesados para eliminar los recursos de aprendizaje duplicados y generar las listas de resultados por cada tipo de expansión. Por último, se realiza la evaluación de la relevancia de los resultados.

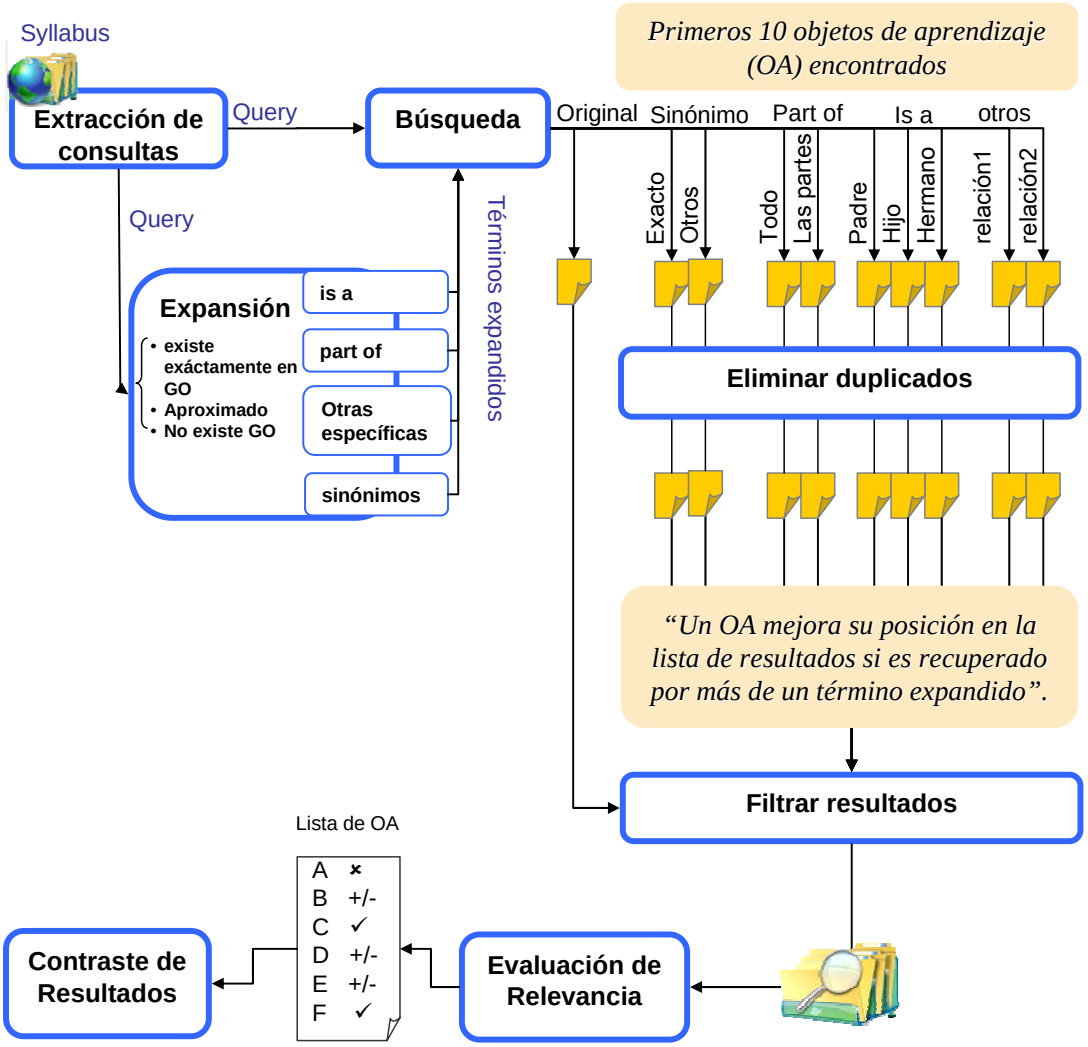


Figura 31: Proceso de evaluación

5.3. Extracción de consultas de prueba

Para definir el conjunto de consultas de prueba se recopilaron 24 syllabus. En la Tabla 16 se incluye el nombre-código del curso, periodo académico, institución que imparte dicho curso y la URL desde donde fue recuperado.

Tabla 16: Colección de Syllabus utilizados en este caso de estudio.

Nombre-Código [a]	**periodo**	**Institución** [b]	**URL**[c]
PCB 2061 Introduction to Genetics	2009	Florida International University	http://www.fiu.edu
Medical genetics	2009	Saba University School of Medicine	http://www.sabamed.org
MBIO 2030 Molecular genetics	2009	University of Antwerp	http://www.ua.ac.be
LS4 Introduction to Genetics	2009	University of California	http://www.lsic.ucla.edu
GENFK - Genetics F K	2008-2009	Mendel University in Brno	http://is.mendelu.cz
ABIO 350 Fundamental Genetics	2009	University of South Carolina	http://www.usca.edu
GENF - Genetics F	2008-2009	Mendel University in Brno	https://is.mendelu.cz
Genetics 302 - Organization of Complex Genomes	2009	University of Alberta	http://www.biology.ualberta.ca
BS803 - Advanced Genetics	2009	The Department of Biological Sciences at Korea Advanced Institute of Science and Technology (KAIST)	http://bio.kaist.ac.kr
BMS 6003 Medical Aspects of Genetics	2009	University of Florida	http://medinfo.ufl.edu
B-KUL-I0D312- 5A Genetics, Genetic Evolution Mechanisms and genetic Nomenclature	2009	Universiteit Leuven	http://www.kuleuven.be
BISC 656-010 Evolutionary Genetics	2009	University of Delaware	http://udel.edu
Biology 4985 Cell and Molecular Biology	2009	University of West Georgia.	http://www.westga.edu
BIOLOGY 3704 Principles of Genetics	2009	Kean University	http://www.kean.edu
BIOLOGY 2344 A Genetics	2009	Georgia Institute of Technology	http://www.goodismanlab.biology.gatech.edu
Biology 1322-1122 Introduction to Genetics	2009	Texas Wesleyan University	http://faculty.txwes.edu
Biol 540 -540L Molecular Genetic	2009	Colorado State University	http://csm.colostate-pueblo.edu
BIOL 362 Principles of Genetics Spring	2009	University of Alaska Fairbanks	http://mercury.bio.uaf.edu
Biol 142 Advanced Topics in Genetics and Molecular Biology	2009	Oxford College of Emory University	https://app.oxford.emory.edu
Biol522 Bacterial Genetics	2009	University of North Carolina	http://www.bio.unc.edu
Bio 496 Genetic and evolution activity	2009	San Diego State University	http://www.sci.sdsu.edu
BIO240 General Genetics	2009	University of Tennessee	http://www.bio.utk.edu
BI 515 Population Genetics	2009	Boston University	http://people.bu.edu
212 Genetics Spring	2009	College Charleston	http://www.neurofly.com

a: Nombre del curso.
b: Institución que imparte el curso.
c: Uniform Resource Locator. URL del dominio donde son públicos los programas de los cursos impartidos.

La definición de las consultas se concentra en la lista de los contenidos propuestos en el curso. Es decir se asume que los contenidos representan los temas en los cuales el diseñador instruccional puede requerir un recurso para el aprendizaje. Por ejemplo, un curso de Algebra I incluye contenidos como *expresiones algebraicas, sistemas de ecuaciones e inecuaciones, funciones y representación gráfica de éstas,* entre otros. Desde esta lista de contenidos se extraen las consultas que el profesor realizaría en un repositorio para recuperar recursos útiles para el curso de algebra.

A partir de todas las listas de contenidos se extrajeron 711 conceptos distintos, para cada uno de los cuales se calculó la frecuencia dentro de la colección de syllabus. Finalmente, fueron seleccionados los conceptos con frecuencia mayor a 1, lo que equivale a 91 consultas de prueba, en la Tabla 17 se detallan las consultas utilizadas.

Tabla 17: Lista de consultas de prueba.

Id	Concepto	Frecuencia[a]	Id	Concepto	Frecuencia[a]
T1	population genetic	12	T46	DNA repair	2
T2	mutation	11	T47	DNA technology	2
T3	DNA replication	10	T48	DNA transcription	2
T4	gene expression	10	T49	evolution	2
T5	recombination	9	T50	evolutionary inference	2
T6	chromosome	7	T51	gene	2
T7	DNA	7	T52	gene interaction	2
T8	mendelian genetic	7	T53	gene isolation	2
T9	transposable element	7	T54	gene manipulation	2
T10	linkage	6	T55	gene mapping in eukaryotes	2
T11	molecular genetic	6	T56	gene therapy	2
T12	biotechnology	5	T57	genetic of virus	2
T13	developmental genetic	5	T58	genotype structure of population	2
T14	genetic	5	T59	granite outcrop	2
T15	meiosis	5	T60	hardy weinberg principle	2
T16	transcription	5	T61	hemoglobinopathy	2
T17	translation	5	T62	heredity	2
T18	ADH in drosophila	4	T63	human disease	2
T19	cytogenetic	4	T64	incompatibility	2
T20	genetic of cancer	4	T65	introduction to DNA	2
T21	genomic	4	T66	linkage of gene	2
T22	introduction	4	T67	mendel law	2
T23	mitosis	4	T68	mendelian inheritance	2
T24	conjugation	3	T69	molecular evolution	2
T25	DNA structure	3	T70	mutation effect	2
T26	gene mapping	3	T71	mutation repair	2
T27	gene regulation	3	T72	non-mendelian inheritance	2
T28	genetic variation	3	T73	PCR	2

Id	Concepto	Frecuencia[a]	Id	Concepto	Frecuencia[a]
T29	genetic of bacteria	3	T74	pedigree	2
T30	inheritance	3	T75	phenotype structure of population	2
T31	migration	3	T76	plasmid	2
T32	non-synonymous substitution	3	T77	protein	2
T33	quantitative genetic	3	T78	random drift	2
T34	quantitative trait	3	T79	recombinant DNA technique	2
T35	transmission genetic	3	T80	recombinant DNA	2
T36	bacterial genetic	2	T81	recombinant DNA technology	2
T37	bacteriophage	2	T82	RNA	2
T38	bioinformatic	2	T83	RNA processing	2
T39	CatLab	2	T84	sex	2
T40	cell cycle	2	T85	sex determination	2
T41	cell cycle control	2	T86	sex linkage	2
T42	cell cycle regulation	2	T87	stem cell	2
T43	centromeres	2	T88	variation in chromosome number	2
T44	cloning	2	T89	variation in chromosome structure	2
T45	disease	2	T90	viral genetic	2
			T91	WRRB	2

a: número de veces que cada concepto se repite dentro del total de conceptos extraídos de la colección de syllabus.

5.4. Expansión

Como se mencionó anteriormente, la expansión de consultas basada en ontologías depende de las relaciones que se utilicen para realizar la expansión. Como se muestra en la Figura 32**,** la ontología Gene posee distintos tipos de relaciones, dentro de las cuales están:

- Relaciones ontológicas básicas, tales como *is_a* y *part_of*
- Relaciones léxicas, tales como sinónimos exacto, general, cercano y relacionado (*synonyms exact, broad, narrow* and *related).*
- Relaciones ontológicas específicas del dominio, en este caso genético, tal como las relaciones de regulación; *regulates, negatively regulates* y *positively regulates*. GO utiliza esta relación para representar la variación de un proceso o su calidad por alguna cosa.

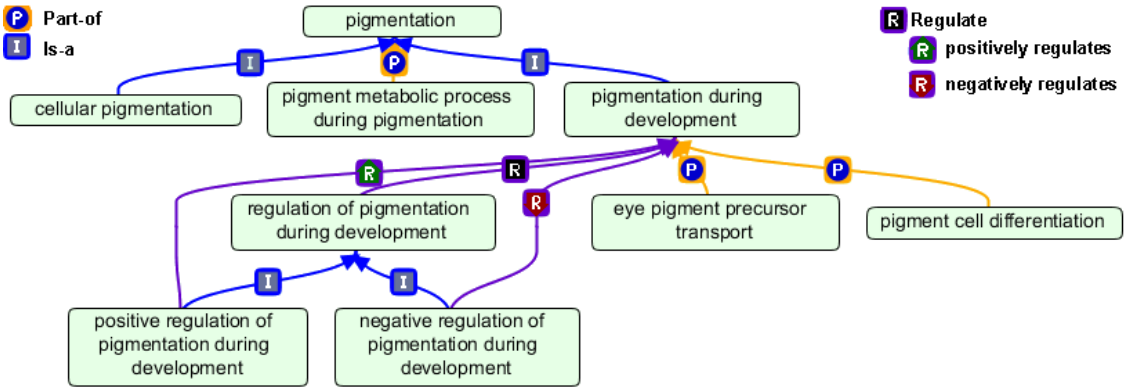


Figura 32: Parte de la Ontología Gene donde se ejemplifican algunas de las relaciones básicas y específicas de una ontología en el dominio de Genética.

A continuación se presenta la especificación de una clase de la ontología Gene en formato OWL (XML-RDF).

```
<owl:Class rdf:about="http://purl.org/obo/owl/GO#GO_0043474">
 <rdfs:label xml:lang="en">pigment metabolic process during
pigmentation</rdfs:label>
 <oboInOwl:hasDefinition>
 <oboInOwl:Definition>
 <rdfs:label xml:lang="en">The chemical reactions and
pathways involving a pigment, any general or particular
coloring matter in living organisms, occurring during the
deposition or aggregation of pigment in an organism, tissue
or cell.</rdfs:label>
 <oboInOwl:hasDbXref>
 <oboInOwl:DbXref>
 <rdfs:label>GOC:jl</rdfs:label>
 <oboInOwl:hasURI
rdf:datatype="http://www.w3.org/2001/XMLSchema#anyURI">http:/
/purl.org/obo/owl/GOC#GOC_jl</oboInOwl:hasURI>
 </oboInOwl:DbXref>
 </oboInOwl:hasDbXref>
 </oboInOwl:Definition>
 </oboInOwl:hasDefinition>
 <oboInOwl:hasExactSynonym>
 <oboInOwl:Synonym>
 <rdfs:label xml:lang="en">pigment metabolism during
pigmentation</rdfs:label>
 </oboInOwl:Synonym>
 </oboInOwl:hasExactSynonym>
<oboInOwl:hasOBONamespace>biological process</oboInOwl:hasOBO
Namespace>
 <rdfs:subClassOf
rdf:resource="http://purl.org/obo/owl/GO#GO 0042440"/>
 <rdfs:subClassOf>
 <owl:Restriction>
 <owl:onProperty>
 <owl:ObjectProperty
rdf:about="http://purl.org/obo/owl/OBO REL#part of"/>
 </owl:onProperty>
 <owl:someValuesFrom
rdf:resource="http://purl.org/obo/owl/GO#GO_0043473"/>
```

```
</owl:Restriction>
</rdfs:subClassOf>
</owl:Class>
```

Además de los tipos de expansión descritos en la sección 4.1.1, y aprovechando las relaciones propias de la ontología Gene, se han agregado los siguientes tipos de expansión:

- **Regulado por** (se abrevia *reg_por*). Se expande a través de los conceptos que están regulados por el concepto buscado. Este tipo de expansión se basa en la relación ontológica *regulate*, incluyendo las relaciones *negatively regulate*s y *positively regulates*.
- **Regula a** (se abrevia *reg_a*). Se expande a través de los conceptos que regulan al concepto buscado. Este tipo de expansión se basa en la relación ontológica *regulate*, incluyendo las relaciones *negatively regulates* y *positively regulates*.
- **Sinónimo exacto** (se abrevia *syn_exa*). Se expande a través de los conceptos que son sinónimos exactos del concepto buscado. Este tipo de expansión se basa en la relación léxica *hasSynonym*.
- **Otros sinónimos** (se abrevia *syn_all*). Se expande a través de los conceptos que son sinónimos del tipo cercano, general y relacionado del concepto buscado (se excluye el tipo sinónimo exacto). Este tipo de expansión se basa en la relación léxica *hasSynonym*, incluyendo las relaciones *hasNarrowSynonym*, *hasBroadSynonym* y *hasRelatedSynonym*.

De las 91 consultas de prueba (listadas antes en la Tabla 17**)**, sólo 15 de ellas buscan un concepto que existe exactamente en la ontología y otras 19 buscan un concepto que tiene un concepto aproximado en la ontología. De estos 34 conceptos buscados en las consultas, 28 poseen algún tipo de expansión en la ontología. Lo anterior depende estrictamente de la ontología, ya que no todos los conceptos representados en ella poseen todos los tipos de relaciones utilizadas para la expansión. La Tabla 18 muestra un resumen con los resultados de los distintos tipos de expansión.

La Figura 33 representa la distribución de la cantidad de nuevos términos agregados por cada tipo de expansión. Se hace la diferencia cuando el concepto buscado tiene una correspondencia exacta con un concepto de la ontología y cuando el concepto buscado no tiene correspondencia exacta y por lo tanto, las expansiones son realizadas con el concepto más cercano. El tipo de expansión *is_a*: hermanos es el tipo que adiciona mayor cantidad de nuevos conceptos, muy por encima del promedio general de 82 conceptos por tipo. Esta situación puede ocurrir por 2 razones, la primera esta relacionada con la ontología, es decir depende de la cantidad de relaciones padre/hijo que existen entre los conceptos modelados. La segunda esta relacionada con la cantidad de veces que se aplicó este tipo de expansión.

Tabla 18: Resumen de la expansión de términos.

	Tipo de expansión								
	isa_exa[a]	*isa_her*[b]	*isa_hij*[c]	*par_exa*[d]	*par_otr*[e]	*reg_a*[f]	*reg_por*[g]	*syn_exa*[h]	*syn_all*[i]
E *	28	28	23	9	8	2	12	9	6
(%)**	(100%)	(100%)	(82%)	(32%)	(29%)	(7%)	(43%)	(32%)	(21%)
eT *v	37	445	140	9	37	2	36	22	12
$\overline{eT}$ *vv	1	16	6	1	5	1	3	2	2

a: Expansión de los conceptos padres del concepto buscado.
b: Expansión de los conceptos hermanos del concepto buscado.
c: Expansión de los conceptos hijos del concepto buscado.
d: Expansión de los conceptos que contienen al concepto buscado.
e: Expansión de los conceptos que están contenidos en el mismo concepto que el concepto buscado.
f: Expansión de los conceptos que regulan al concepto buscado.
g: Expansión de los conceptos que son regulados por el concepto buscado.
h: Expansión de los conceptos sinónimos exactos.
i: Expansión de los conceptos sinónimos cercano, general o relacionado (excepto sinónimo exacto).
*: Cantidad de veces que cada tipo de expansión fue aplicado.
**: Porcentaje respecto al total de consultas que tienen por lo menos un tipo de expansión (dado el total de 28 consultas).
*v: Cantidad de términos expandidos por cada tipo de expansión.
*vv: Promedio de términos por cada expansión.

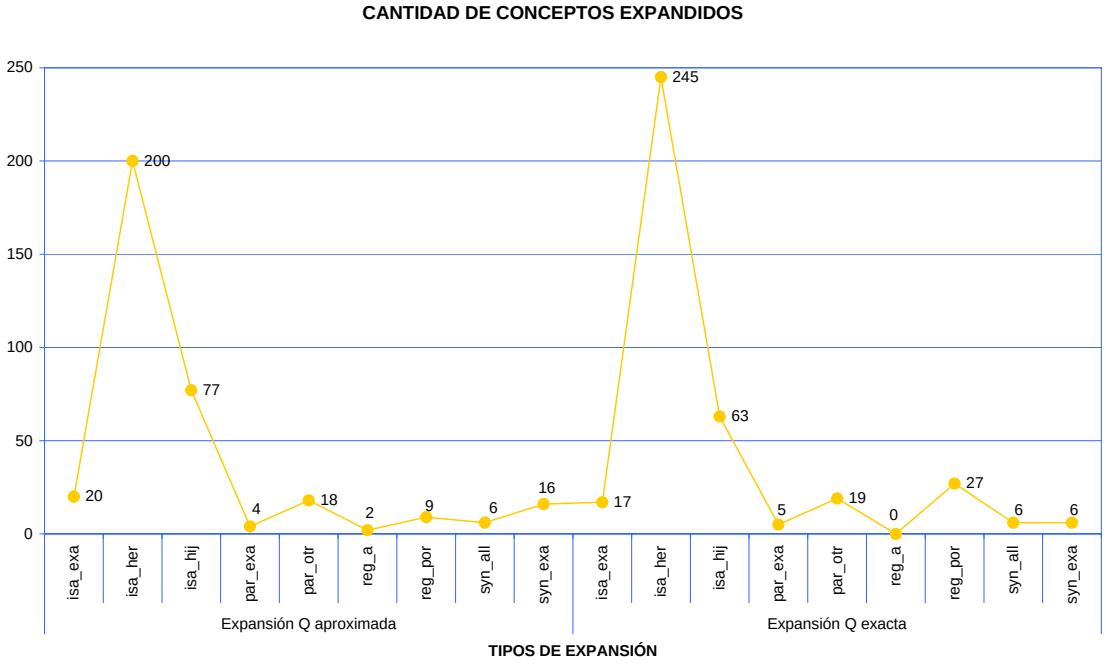


Figura 33: Gráfica del resumen de los resultados de la expansión en las consultas que coinciden exactamente con algún concepto en la ontología y en las consultas donde se utilizó el concepto más cercano.

Según se indica en la Tabla 18, este tipo de expansión se aplicó en 28 consultas, al igual que el tipo de expansión *is_a*: hijos, por lo tanto la alta cantidad de términos expandidos

queda mejor explicada por la cantidad de relaciones de herencia que poseen los conceptos representados en la ontología.

La Tabla 19 presenta la frecuencia acumulada de la cantidad de conceptos agregados tras la expansión. El número de conceptos expandidos van desde 0 (sin expansión) hasta 48 conceptos expandidos. Como se destaca en la tabla en un 75% de los casos la expansión llega hasta 6 conceptos.

Tabla 19: Cantidad de conceptos expandidos por consulta y frecuencia en el total de expansiones realizadas.

Clase, número de conceptos expandidos	Frecuencia	% acumulado
1	**42**	**33,60%**
2	**18**	**48,00%**
4	**22**	**65,60%**
6	**12**	**75,20%**
9	9	82,40%
14	8	88,80%
20	6	93,60%
y mayor...	8	100,00%

5.5. Búsqueda en el repositorio

Considerando las opciones y restricciones del servicio Restful de MERLOT, en cuanto al tratamiento de la cadena de búsqueda descritas en la sección 5.1.2, se realizaron consultas separadas tanto para los términos expandidos como para la consulta original. Por ejemplo, el concepto consultado *translation* tiene 5 conceptos como sinónimos exactos: *protein biosynthesis*, *protein translation, protein synthesis, protein formation, y protein anabolism*. Por lo tanto se ejecutaron 6 consultas, para cada concepto por separado. A continuación se especifican las consultas para este ejemplo.

```
Qoriginal
"http://www.merlot.org/merlot/materials.rest?licenseKey=VALID
A&keywords=
translation&creativeCommons=false&exactPhraseKeyWords=true"

Qexpandida 1
"http://www.merlot.org/merlot/materials.rest?licenseKey=VALID
A&keywords=protein %20biosynthesis
&creativeCommons=false&exactPhraseKeyWords=true"

Qexpandida 2
http://www.merlot.org/merlot/materials.rest?licenseKey=VALIDA
&keywords= protein %20translation
&creativeCommons=false&exactPhraseKeyWords=true

Qexpandida 3
http://www.merlot.org/merlot/materials.rest?licenseKey=VALIDA
&keywords= protein %20synthesis
&creativeCommons=false&exactPhraseKeyWords=true
```

```
Qexpandida 4
"http://www.merlot.org/merlot/materials.rest?licenseKey=VALID
A&keywords= protein %20formation
&creativeCommons=false&exactPhraseKeyWords=true"

Qexpandida 5
"http://www.merlot.org/merlot/materials.rest?licenseKey=VALID
A&keywords= protein %20anabolism
&creativeCommons=false&exactPhraseKeyWords=true"
```

Los resultados de las consultas contienen los metadatos de los OA que la satisfacen, por ejemplo, a continuación se presenta el resultado entregado por el repositorio para la `Qexpandida 2`, note que en este caso sólo se recuperaron 2 OA (`totalCount`).

```
<?xml version="1.0" encoding="UTF-8"?>
<merlotMaterialSearchWebService
xmlns="http://www.merlot.org/merlot/materials-rest"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
<status>ok</status>
<summary>
     <totalCount>2</totalCount>
     <resultCount>2</resultCount>
     <lastRecNumber>2</lastRecNumber>
</summary>
<results>
     <material>
          <title>Protein Translation</title>
          <URL>http://bioweb.uwlax.edu/GenWeb/Molecular/Theor
     y/Translation/translation.htm</URL>
          <authorName>Scott Cooper</authorName>
          <creationDate>1109923200000</creationDate>
          <description>A collection of interactive animations
     that allow studnets to move ribosomal subunits and
     transfer RNAs to translate a mRNA into protein. Includes
     background text and practice problems.</description>
          <detailURL>http://www.merlot.org/merlot/viewMateria
     l.htm?id=83252</detailURL>
          <creativeCommons>false</creativeCommons>
     </material>
     <material>
           <title>Blazing a Genetic Trail</title>
           <URL>http://www.hhmi.org/genetictrail/index.html</
     URL>
           <authorName>Edited by Maya Pines</authorName>
           <creationDate>1143014400000</creationDate>
           <description>The web version of a Howard Hughes
     Medical Institute report on genetic diseases. It offers
     descriptions of the research and discovery of disease
     like cystic fibrosis and Huntington's disease. It also
     discusses different forms of inherence (dominant,
     recessive, sex-linked), as well as discussions of
     mutation, protein translation and structure.
     </description>
           <detailURL>http://www.merlot.org/merlot/viewMateri
     al.htm?id=86288</detailURL>
           <creativeCommons>false</creativeCommons>
     </material>
</results>
</merlotMaterialSearchWebService>
```

Los resultados obtenidos para cada consulta se encuentran ordenados según el ranking general estimado por el repositorio. A medida que aumenta la posición en la lista de

resultados se reducen las posibilidades de que el usuario revise el OA recuperado. Estudios como el realizado por (Spink, Jansen, Wolfram, & Saracevic, 2002) sobre el comportamiento del usuario en la búsqueda Web indican que la mayoría de los usuarios revisa sólo la primera página de resultados. De acuerdo a lo anterior, para nuestro análisis sólo fueron procesados los 10 primeros objetos de aprendizaje recuperados por cada consulta. Por ejemplo para el tipo de expansión de sinónimo exacto (*syn_exa*) de la consulta T17 *Translation*, el concepto consultado *protein synthesis* localiza 32 OA en el repositorio, de ellos se recuperan sólo los primeros 10 OA del ranking (refiérase a la Tabla 20).

Tabla 20: Ejemplo resultados encontrados y recuperados desde el repositorio, para la consulta T17 *Translation* y tipo de expansión de sinónimo exacto (*syn_exa*).

id_Q	Consulta Q	Conceptos expandidos del tipo sinónimo exacto	tOA[b]	rOA[a]
T17	translation	protein biosynthesis	0	**0**
		protein translation	2	**2**
		protein synthesis	32	**10**
		protein formation	0	**0**
		protein anabolism	0	**0**

a: cantidad total de OA recuperados.
b: cantidad total de OA que responden a la consulta en el repositorio.

En el Anexo A se detallan los resultados entregados por el repositorio para cada una de las consultas de prueba y sus expansiones.

Para resumir los resultados de la expansión y búsqueda, en la Tabla 21 se presentan los casos en los cuales se aplicó cada tipo de expansión y el porcentaje de las veces en las que se obtuvieron resultados (OA) en el repositorio. Por ejemplo la expansión a través de la relación *part_of*: el Todo (*part_exa*) sólo se aplicó en 9 de las 28 consultas iniciales, de estas 9 sólo en 4 de ellas (44%) se obtuvieron OA en el repositorio.

Según se muestra en la Figura 34 las expansiones, *is_a*: hermanos *(isa_her*), *is_a*: hijos (*isa_hij*) e *is_a*: padre (*isa_exa*) son los tipos de expansiones que localizan y recuperan mayor cantidad de OA. Además son estos mismos tipos los que se aplican una mayor cantidad de veces y a la vez agregan una mayor cantidad de nuevos términos. No obstante lo anterior, tras analizar en detalle el promedio de OA recuperados por cada término expandido se obtiene que los tipos de expansión con mayor cantidad de OA recuperados son *part_of*: el todo (*par_exa*), *part_of*: las partes (*par_otr*) e *is_a*: padre (*isa_exa*).

Tabla 21: Resumen del proceso de búsqueda.

	Tipo de expansión								
	isa_exa[a]	*isa_her*[b]	*isa_hij*[c]	*par_exa*[d]	*par_otr*[e]	*reg_a*[f]	*reg_por*[g]	*syn_exa*[h]	*syn_all*[i]
E*	28	28	23	9	8	2	12	9	6
S/OA**	20	12	13	5	5	1	10	7	6
C/OA***	8	16	10	4	3	1	2	2	0
(C/OA % de E)[v]	29%	57%	43%	44%	38%	50%	17%	22%	0%

a: Expansión de los conceptos padres del concepto buscado.
b: Expansión de los conceptos hermanos del concepto buscado.
c: Expansión de los conceptos hijos del concepto buscado.
d: Expansión de los conceptos que contienen al concepto buscado.
e: Expansión de los conceptos que están contenidos en el mismo concepto que el concepto buscado.
f: Expansión de los conceptos que regulan al concepto buscado.
g: Expansión de los conceptos que son regulados por el concepto buscado.
h: Expansión de los conceptos sinónimos exactos.
i: Expansión de los conceptos sinónimos cercano, general o relacionado (excepto sinónimo exacto).

*: Cantidad de veces que cada tipo de expansión fue aplicado
**: Cantidad de veces que cada tipo de expansión NO recuperó OA en el repositorio.
***: Cantidad de veces que cada tipo de expansión recuperó OA en el repositorio.
v: Porcentaje respecto al total de consultas que tienen cada tipo de expansión (E).

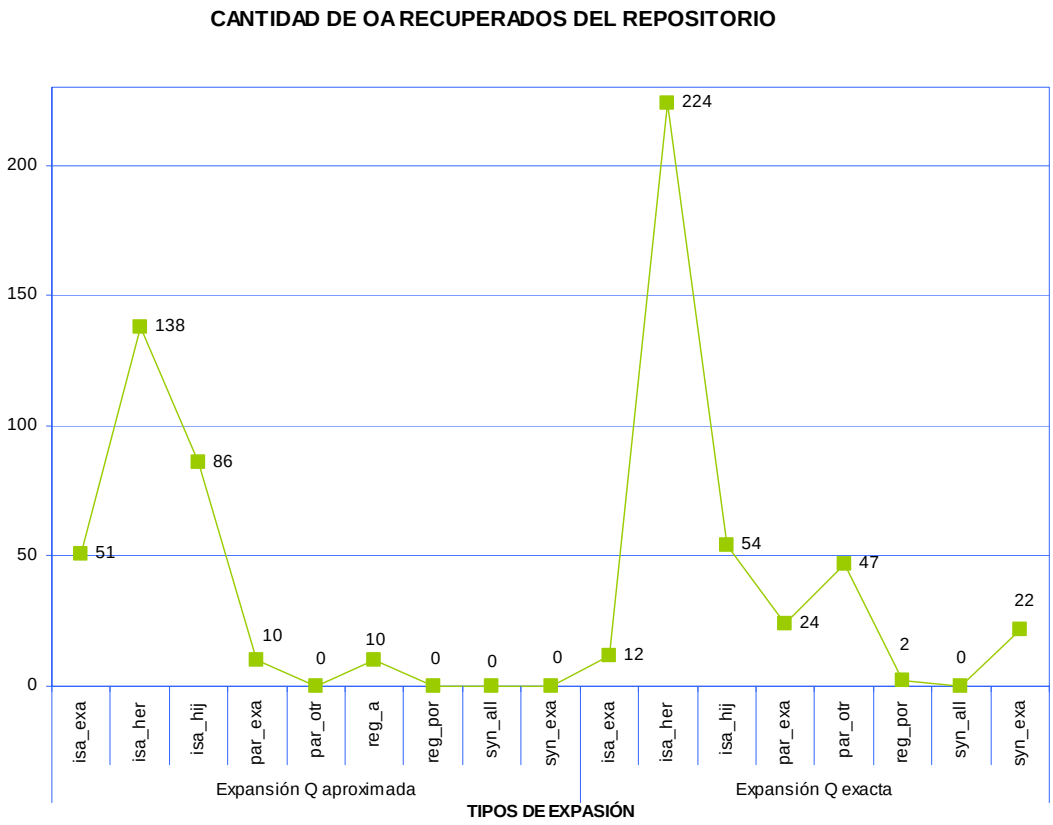


Figura 34: Gráfica de resumen de los resultados de las consultas ejecutadas en el repositorio.

De las 28 consultas que poseen algún tipo de expansión se recuperaron 895 OA, de ellos:

- **315 son OA únicos dentro del conjunto total**, 162 de éstos poseen repeticiones. La cantidad de repeticiones varía entre 1 y 25, su análisis se presenta en la Tabla 22.

Tabla 22: Análisis de frecuencia de la cantidad de repeticiones de cada OA recuperádo.

Cantidad de repeticiones	Frecuencia	% acumulado
1	153	48,57%
3	74	72,06%
5	56	89,84%
7	8	92,38%
9	9	95,24%
11	7	97,46%
y mayor...	8	100,00%

- **816 son OA únicos dentro de una misma consulta y tipo de expansión**, 47 de éstos poseen repeticiones. La cantidad de repeticiones varía entre 1 y 4, su análisis se presenta en la Tabla 23.

Tabla 23: Análisis de frecuencia de la cantidad de repeticiones de cada OA recuperado único para una misma consulta y tipo de expansión.

Cantidad de repeticiones	Frecuencia	% acumulado
1	769	94,01%
2	26	97,19%
3	12	98,66%
4	11	100,00%
y mayor...	0	100,00%

- **673 son OA únicos dentro de una misma consulta**, 128 de éstos poseen repeticiones. La cantidad de repeticiones varía entre 1 y 9, su análisis se presenta en la Tabla 24.

Tabla 24: Análisis de frecuencia de la cantidad de repeticiones de cada OA recuperado único para una misma consulta.

Cantidad de repeticiones	Frecuencia	% acumulado
1	545	80,98%
2	98	95,54%
3	16	97,92%
4	4	98,51%
y mayor...	10	100,00%

Así como fue explicado en la sección 4.1, la separación de la cadena de búsqueda provoca que existan resultados repetidos para un mismo tipo de expansión y consulta, de ahí que según la solución propuesta para estos casos, se recalcula la posición de cada uno de los OA repetidos dentro de una misma consulta y tipo de expansión.

A partir de la lista de resultados de cada concepto expandido se creó una lista única donde los OA repetidos fueron reposicionados según la Ecuación 9. En la Tabla 25 se detalla el proceso de eliminación de duplicados para la consulta T15 *meiosis* y el tipo de expansión *is_a*: hermanos.

Tabla 25: Eliminación de duplicados en las listas de resultados obtenidos de los términos expandidos, consulta T15 *meiosis*, tipo expansión *is_a*: hermanos.

OA	Posiciones ranking en repeticiones[a]	nº veces repetidos[b]	Me[c]	POA[d]	Ranking final 10 primeros resultados [e]
81127	0	1	0	0	1
91401	0 0	2	0	0	2
78511	1	1	1	1	3
84051	1	1	1	1	4
90943	3 1	2	2	1	5
243069	2	1	2	2	6
75251	2	1	2	2	7
78646	2	1	2	2	8
77749	8 5 7	3	7	2,3	9
91398	6 4	2	5	2,5	10
229222	3	1	3	3	
81501	3	1	3	3	
271486	4	1	4	4	
80693	4	1	4	4	
272275	5	1	5	5	
90383	5	1	5	5	
86024	6	1	6	6	
90307	6	1	6	6	
79615	7	1	7	7	
80042	7	1	7	7	
82020	8	1	8	8	
85025	8	1	8	8	
80790	9	1	9	9	
84029	9	1	9	9	
84159	9	1	9	9	

b: es el número de veces que el j-ésimo OA esta repetido en las listas de resultados de las expansiones.
c: Me es la mediana de las posiciones en las que el j-ésimo OA ha sido recuperado.
d: POA nueva posición de ranking del OA
e: la lista final ordenada por valor de POA

Una vez calculadas las nuevas posiciones de ranking, se eliminan las instancias de repetición, y se genera una lista única por cada tipo de expansión. El detalle del procesamiento para re -calcular la posición de los resultados repetidos se presenta en el Anexo B .

Como resultado del proceso anterior se obtiene la lista de los 10 primeros OA que responden a cualquiera de los conceptos expandidos (*lista individual por tipo de expansión*), es decir:

```
(concepto expandido A1 OR concepto expandidoA2 OR
…concepto_expandido AN-1)
```

Por lo tanto, para recomponer la cadena de búsqueda completa para cada tipo de expansión es preciso unir estos resultados con los resultados de la consulta original (*lista por tipo de expansión*), es decir:

```
Q, Tipo expansión A:

consulta original OR (concepto expandido A1 OR
concepto_expandidoA2 OR …concepto_expandido AN-1)
```

Por ejemplo, cadena de búsqueda para la consulta T15 *meiosis,* expandida a través de la relación *is_a*: hermanos sería:

```
Q: MEIOSIS, tipo expansion IS_A: HERMANOS:

meiosis OR (meiosis I OR mitosis OR S-phase)
```

La unión de los 10 OA recuperados de los conceptos expandidos con los 10 OA recuperados de la consulta original se realiza por medio de la intercalación de ambas listas. Si existen OA repetidos se asume su mejor posición de ranking. En la Tabla 26 la primera columna contiene la lista de OA recuperados de la consulta inicial, la segunda columna contiene la lista única de OA recuperados para la expansión *is_a*: hermanos, y la última columna contiene la unión de ambas listas (OR), es decir la lista final de consulta expandida T15 *meiosis*. Los OA repetidos se encuentran tachados, por lo tanto son descartados de la lista final.

Tabla 26: Lista de OA recuperados, ejemplo de la consulta original OR expansión, Consulta T15 *Meiosis.*

ranking	OA recuperados		
	original	lista individual por tipo de expansión *isa_her*	lista por tipo de expansión original OR *isa_her*
1	91401	81127	91401
2	84051	~~91401~~	81127
3	90943	78511	84051
4	75251	~~84051~~	90943
5	81501	~~90943~~	78511
6	91398	243069	75251
7	77749	75251	81501

ranking	OA recuperados		
	original	lista individual por tipo de expansión *isa_her*	lista por tipo de expansión original OR *isa_her*
8	86024	78646	91398
9	80042	77749	243069
10	82020	91398	77749

5.6. Filtrado de los resultados

Una vez que se han generado las listas de resultados (OA) para cada consulta y tipo de expansión, continúa la evaluación de la relevancia de los resultados. Sin embargo la gran cantidad de instancias de evaluación y el costo en tiempo de los expertos asociado a la evaluación, hace necesario restringir el conjunto de resultados a evaluar. Para esto se han definido 3 restricciones:

Restricción 1 (R1): Consultas que existan exactamente en la ontología. Es decir, se evaluarán los resultados de las consultas donde el concepto buscado coincide con un concepto de la ontología.

Restricción 2 (R2): Consultas con una frecuencia mayor que cuatro en la colección de syllabus. Es decir, se evaluarán los resultados de las consultas cuando el concepto buscado es más representativo en el conjunto de syllabus.

Restricción 3 (R2): Consultas en las que se haya aplicado por lo menos un tipo de expansión, independiente del tipo. Como fue mencionado antes no todas los tipos de expansión pueden ser aplicadas a cada consulta, esto depende de la ontología. Por lo tanto, para efectos de análisis, se evaluarán los resultados de las consultas con más de un tipo de expansión.

Cabe recordar que en 28 de las 91 consultas se recuperaron resultados en el repositorio. Al aplicar la restricción R1 el conjunto se reduce a 15 de las 28 consultas. El conjunto de resultados a evaluar para las 15 consultas es de 281 instancias. Al incorporar la segunda restricción, $(R1 \cap R2)$ el conjunto a evaluar se reduce a 7 de las 15 consultas, lo que equivale a su 47%. Con la tercera restricción, $((R1 \cap R2) \cap R3)$ el conjunto de consultas se reduce definitivamente a 5 de las 15 consultas, lo que equivale a su 33.3%. El conjunto de resultados a evaluar para las 5 consultas es de 98 instancias, de ellas 70 son únicas (es decir, no repetidas).

En resumen, pasarán a la evaluación de los expertos los OA recuperados de las 5 consultas que existen exactamente en la ontología, que tienen una frecuencia mayor que 4 en la colección de syllabus, y que tienen al menos un tipo de expansión.

Los resultados obtenidos para las 5 consultas y sus expansiones se detallan en la Tabla 27. En ella se incluye la consulta, el tipo de expansión, el concepto consultado y la cantidad total de OA recuperados (rOA).

Tabla 27: Resultados de la búsqueda en el repositorio MERLOT después de aplicadas las restricciones (R1∩R2) ∩R3.

Id	Consulta	Tipo	Conceptos expandidos	rOA*	TOA**	
T3	DNA replication	original	DNA replication	10	10	10
		isa_her	base-excision repair gap-filling	10	42	10 primeros
			DNA recombination	1		
			DNA repair	3		
			nucleotide-excision repair DNA gap filling	7		
			poly-gamma-glutamate biosynthetic process	1		
			transcription	10		
			translation	10		
		isa_hij	DNA-dependent DNA replication	10	20	10 primeros
			RNA-dependent DNA replication	10		
T6	chromosome	original	chromosome	10	10	10
		isa_exa	intracellular non-membrane-bounded organelle	2	2	2
		isa_her	aster	10	18	10 primeros
			cytoskeleton	8		
T15	meiosis	original	meiosis	10	10	10
		isa_her	meiosis I	10	30	10 primeros
			mitosis	10		
			S phase	10		
T16	transcription	original	transcription	10	10	10
		isa_her	DNA replication	10	27	10 primeros
			nucleotide-sugar metabolic process	7		
			translation	10		
		par_exa	gene expression	8	8	8
		par_otr	regulation of gene expression	1	13	10 primeros
			RNA processing	2		
			translation	10		
T17	translation	original	translation	10	10	10
		isa_her	DNA replication	10	20	10 primeros
			transcription	10		
		par_ex	gene expression	8	8	8

Id	Consulta	Tipo	Conceptos expandidos	rOA*	TOA**	
		a				
		par_otr	regulation of gene expression	1	13	10 primeros
			RNA processing	2		
			transcription	10		
		reg_por	regulation of translation	1	1	1
		syn_exa	protein synthesis	10	12	10 primeros
			protein translation	2		

*: Cantidad de OA recuperados por cada consulta
**: Suma total de OA por tipo de expansión.

5.7. Evaluación de relevancia

En el campo de la RI en general existen colecciones de prueba estándar, tales como por ejemplo las colecciones TREC (*Text Retrieval Conference*), NTCIR (*NII Test Collections for IR Systems*), CLEF (*Cross Language Evaluation Forum*), etc. A partir de estas colecciones es posible realizar la evaluación del desempeño, la comparación con otras propuestas realizadas en investigaciones previas y la reproducción de experimentos en las mismas condiciones (Voorhees, 2000). Las colecciones de prueba por ejemplo incluyen:

- Colección de documentos con su definición (fuente de los documentos, periodo, tamaño, etc.).
- Listado con las consultas de prueba.
- Juicios de relevancia de los resultados para cada consulta.

A partir de esta información es posible calcular distintas métricas de desempeño del RI y contrastar los resultados con propuestas previas. En el área de la recuperación de objetos de aprendizaje desde repositorios no existen este tipo de colecciones, lo cual dificulta la estimación de algunas métricas de desempeño en la recuperación de OA. En nuestra investigación, el análisis del desempeño de la expansión se basa en la evaluación de la relevancia realizada por expertos.

En el experimento participaron tres expertos, los cuales tienen el perfil de médicos o especialistas en genética, con un mínimo de experiencia de 5 años y con experiencia en docencia.

Los OA recuperados para cada consulta, con y sin expansión, serán evaluados por los expertos de acuerdo a 3 niveles ordenados de respuesta (Jarvelin & Kekalainen, 2000).

- Relevante. Abreviado R y cuya ponderación es 1.0.
- Parcialmente Relevante. Abreviado PR y cuya ponderación es 0.5.
- No relevante. Abreviado N y cuya ponderación es 0.

La ponderación asignada a cada nivel de respuesta permite refinar la estimación de la métrica de precisión, considerando el grado de relevancia de los resultados (Dolog et al., 2008).

La evaluación de la relevancia de un OA recuperado se basa en la temática involucrada en la consulta y el contenido del OA, es decir la evaluación es del **tipo de relevancia temática** (*topically relevance*) (Borlund, 2003). La pregunta de evaluación que el experto debe contestar al evaluar cada OA recuperado es: ¿*El objeto de aprendizaje satisface la consulta aplicada*?.

Un extracto del instrumento utilizado para la evaluación de los expertos se incluye en el Anexo C .

Para validar la evaluación de la relevancia de los OA se analiza el nivel de acuerdo entre las evaluaciones de los expertos a través del análisis de *Kappa* no ponderado de *Cohen* (Blackman & Koval, 2000; Uebersax, 1983). Cuyo valor κ se calcula según la Ecuación 10:

$$\kappa = \frac{po - pe}{1 - pe} \qquad \text{Ecuación 10}$$

Donde:
p_o: el acuerdo observado entre los evaluadores (Ecuación 11).
p_e: probabilidad hipotética de acuerdo (Ecuación 12).

$$po = \frac{n\_acds}{n\_acds + n\_ds\_acds} \qquad \text{Ecuación 11}$$

Donde:
`n_acds`: cantidad de instancias de evaluación en las cuales la evaluación de los expertos coincide.
`n_ds_acds`: cantidad de instancias de evaluación en las cuales la evaluación de los expertos No coincide.

$$pe = \sum_{i=1}^{n} (p_{i1} \times p_{i2}) \qquad \text{Ecuación 12}$$

Donde:
`n`: cantidad de categorías.
`i`: número de la categoría (de `1` hasta `n`).
p_{i1}: proporción de ocurrencias de la categoría `i` para el observador 1.
p_{i2}: proporción de ocurrencias de la categoría `i` para el observador 2.

La Tabla 28 muestra los niveles para la interpretación de los valores del coeficiente Kappa propuestos por (Landis & Koch, 1977). Si los evaluadores están en acuerdo completamente $\varkappa$ es igual a 1, si no hay acuerdo entre los evaluadores (aparte de lo que cabría esperar por el azar) $\varkappa \leq 0$.

Tabla 28: Niveles para la interpretación del coeficiente Kappa.

$\varkappa$	Interpretación
< 0	Sin acuerdo (*No agreement*)
0.0 — 0.20	Acuerdo leve (*Slight agreement*)
0.21 — 0.40	Acuerdo justo (*Fair agreement*)
0.41 — 0.60	Acuerdo moderado (*Moderate agreement*)
0.61 — 0.80	Acuerdo sustancial (*Substantial agreement*)
0.81 — 1.00	Acuerdo casi perfecto (*Almost perfect agreement*)

Como se observa en la Tabla 29 el nivel de acuerdo entre el evaluador 1 y 2, el evaluador 1 y 3, y el evaluador 2 y 3, es moderado. En promedio, el valor del coeficiente Kappa $\boldsymbol{\kappa}$ es 0.4880, lo que también indica que existe un nivel de acuerdo moderado (es decir, 0.41 $\leq \varkappa <$0.6). En la Tabla 29 parte (a) se resumen las evaluaciones politómicas entre los tres evaluadores y la parte (b) incluye el resumen de los valores del coeficiente Kappa.

Tabla 29: Coeficiente Kappa. Resumen para los 3 niveles de respuesta (relevante, parcialmente y no relevante) con 3 evaluadores.

(a)

		RATER 2			
		R	PR	N	
RATER 1	R	**36**	2	4	
	PR	16	**8**	5	
	N	2	1	**24**	
					∑98

		RATER 3			
		R	PR	N	
RATER 1	R	**34**	7	1	
	PR	14	**11**	4	
	N	3	1	**23**	
		45	27	26	∑98

		RATER 3			
		R	PR	N	
RATER 2	R	**38**	10	6	
	PR	7	**4**	0	
	N	6	5	**22**	
					∑98

b)

RATER 1-2	po	0,69387755	RATER 1-3	po	0,69387755	RATER 2-3	po	0,65306122
	pe	0,35582712		pe	0,3591212		pe	0,4047272
	$\boldsymbol{\kappa_{1-2}}$	0,52478216		$\boldsymbol{\kappa_{1-3}}$	0,52233956		$\boldsymbol{\kappa_{2-3}}$	0,41717684

Si bien existe un acuerdo moderado entre los evaluadores, una posible razón para las diferencias entre las evaluaciones de los expertos está relacionada con el nivel de granularidad y tipo de recursos que se recuperan desde el repositorio. Los recursos van desde cursos hasta libros, artículos o animaciones específicas. Además, muchos de los recursos son colecciones de OA, por lo cual se dificulta la evaluación única de la relevancia de la colección.

El coeficiente de correlación de *Spearman* o simplemente rho de *Spearman* (definido como ρ), es una medida no paramétrica de la dependencia entre dos variables. Su análisis permite determinar si existe relación entre dos variables y si esta relación es significante. El estadístico ϱ viene dado por la expresión (Spearman, 1987):

$$\rho = 1 - \frac{6\sum D^2}{N\left(N^2 - 1\right)} \qquad \text{Ecuación 13}$$

Donde
D: es la diferencia entre los correspondientes valores de x-y.
N: es el número de parejas.

La interpretación del coeficiente rho de Spearman tiene los siguientes 3 niveles (Ecuación 14):

$$\rho = \begin{cases} < 0{,}3 & ; \quad \textit{asociación débil} \\ 0{,}3 \leq \rho < 0{,}7 & ; \quad \textit{asociación moderada} \\ \geq 0{,}7 & ; \quad \textit{asociación fuerte} \end{cases} \qquad \text{Ecuación 14}$$

Con este coeficiente se evalúa cuan bien se relacionan las evaluaciones de relevancia entre los expertos. En la Tabla 30, se muestra que existe asociación entre las evaluaciones de relevancia otorgadas por los tres evaluadores. El nivel de asociación entre los evaluadores es moderado (es decir, $\rho \geq 0.3$ y $\rho < 0.7$) con un nivel de significancia de 0.01 (bilateral).

Tabla 30: Análisis no paramétrico de Rho Spearman. Asociación general entre los niveles de relevancia, otorgada por los tres expertos, a los OA recuperados para cada consulta.

		Rater 1	Rater 2	Rater 3
Rater 1	Coeficiente de correlación	1,000	**,668****	**,662****
	Significancia (bilateral)	.	,000	,000
	N	98	98	98
Rater 2	Coeficiente de correlación	**,668****	1,000	**,512****
	Significancia (bilateral)	,000		,000
	N	98	98	98
Rater 3	Coeficiente de correlación	**,662****	**,512****	1,000
	Significancia (bilateral)	,000	,000	.
	N	98	98	98

** La correlación es significante en 0.01 (bilateral).

5.8. Contraste de resultados

Las métricas más comunes para evaluar el desempeño en la RI son la precisión y la exhaustividad (*recall*) (Baeza-Yates & Ribeiro-Neto, 1999; Borlund, 2003; Dolog et al., 2008; Jarvelin & Kekalainen, 2000; Voorhees, 2001). Tal como fue explicado en la sección 2.4.1, la precisión se calcula como la razón entre los resultados relevantes recuperados respecto al total de resultados recuperados (ver la Ecuación 4 descrita en la página 42) y la exhaustividad (*recall*) se obtiene como la razón entre los resultados relevantes recuperados y el total de resultados relevantes de la colección (ver la Ecuación 3 descrita en la página 42).

Tal como fue explicado en la sección anterior, en el ámbito de la recuperación de OA en repositorios no existe una colección de prueba que nos permita conocer la cantidad total de resultados relevantes en la colección para cada consulta de prueba, por consiguiente en nuestro experimento no es posible calcular la métrica de recall.

Para efectos de éste análisis la métrica de precisión será complementada con la métrica DCG (*cumulated gain with discount by document rank*), esta última es una mejora a la métrica tradicional de precisión. DCG disminuye la contribución de la relevancia acumulada en la medida que se avanza en el ranking de los resultados. Es decir a mayor ranking es menor la contribución del documento a la ganancia acumulada, ya que es menos probable que el usuario examine el documento por el tiempo, esfuerzo y la información acumulada de los documentos previamente revisados. Como se observa en la Ecuación 15, la función para reducir progresivamente el valor de relevancia del documento es el logaritmo en base 2 de la posición en el ranking.

$$DCG[i] = \begin{cases} G[1] & ;i = 1 \\ DCG[i-1] + \dfrac{G[i]}{\log_2 i} & ;i \neq 1 \end{cases} \qquad \text{Ecuación 15}$$

Donde:
`G`: es el vector de ganancia.
$G[i]$: es la relevancia del documento en la posición del ranking `i`.

Para contrastar los resultados obtenidos con la expansión basada en ontología de dominio, se utilizan 2 métricas orientadas al usuario; la novedad (*novelty*) y cobertura (*coverage*). Estas métricas permiten observar si con la expansión se obtienen resultados relevantes distintos a los que se obtienen sin la expansión.

La cobertura se define como la fracción de los documentos relevantes conocidos recuperados respecto al total de documentos relevantes conocidos (Baeza-Yates & Ribeiro-Neto, 1999), véase la Ecuación 16.

La novedad es la fracción de los documentos relevantes desconocidos que han sido recuperados respecto a los documentos relevantes recuperados. Es decir, los documentos relevantes que el usuario no esperaba encontrar (Baeza-Yates & Ribeiro-Neto, 1999), véase la Ecuación 17.

$$\text{coverage} = \frac{|Rk|}{|U|}$$ Ecuación 16

Donde:
U: es el conjunto de resultados relevantes conocidos.
|Rk|: es el número de resultados relevantes conocidos que fueron recuperados por la consulta.

$$\text{novelty} = \frac{|Ru|}{|Ru| + |Rk|}$$ Ecuación 17

Donde:
|Rk|: es el número de documentos relevantes conocidos que fueron recuperados por la consulta.
|Ru|: es el numero de documentos relevantes desconocidos que fueron recuperados por la consulta, es decir documentos relevantes que no pertenecen al conjunto U.

Según se representa en la Figura 35 se tiene:

Q	Conjunto de documentos recuperados de la consulta.
R	Conjunto de documentos relevantes de la colección para la consulta.
U	Conjunto de documentos relevantes de la colección para la consulta que son conocidos por el usuario.
$U \subset R$	El conjunto U es un subconjunto de R.
$Rk \subset Q$	El conjunto Rk es un subconjunto de Q.
$Ru \subset Q$	El conjunto Ru es un subconjunto de Q.
$Rk \subset U$	El conjunto Rk es un subconjunto de U.
$Ru \subset R$	El conjunto Ru es un subconjunto de R.
$(Rk \cup Ru) \subset R$	La unión de los elementos de Rk y Ru es un subconjunto de los documentos relevantes de la colección R.
$(Rk \cup Ru) \subset Q$	La unión de los elementos de Rk y Ru es un subconjunto de los

documentos recuperados de la consulta Q.

(R∩A) = (Rk∪Ru) La intersección entre el conjunto de documentos recuperados de la consulta Q y el conjunto de los documentos relevantes de la colección, R equivale a la unión de los elementos de Rk y Ru.

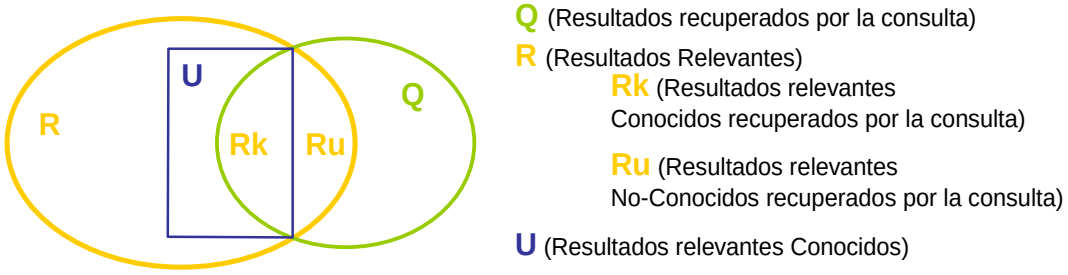


Figura 35: Métricas de cobertura y novedad para una consulta de ejemplo tomada de (Baeza-Yates & Ribeiro-Neto, 1999), pp. 83.

Dado que los elementos de los conjuntos R y U son datos desconocidos en nuestro experimento, la cobertura y la novedad de las consultas expandidas *se calcula en función de la cantidad de resultados obtenidos de la consulta original (es decir, sin expansión*). En otras palabras, se asume que el conjunto U (conjunto de documentos relevantes conocidos por el usuario) corresponde al conjunto de los resultados relevantes recuperados en la consulta original.

Según lo anterior, la cobertura corresponde a la fracción de los documentos relevantes conocidos recuperados por la consulta expandida Rk, que pertenecen al conjunto de documentos relevantes conocidos U. La novedad corresponde a la fracción de documentos relevantes desconocidos recuperados por la consulta expandida, Ru respecto al total de documentos relevantes recuperados de la consulta.

Por ejemplo, si una consulta recupera los siguientes objetos Q={A,B,C,D,G,F}, cuyas evaluaciones de relevancia son {R,N,R,PR,PR,N}. Como se observa en la Figura 36, el conjunto de objetos relevantes conocidos queda definido por U={A,C,D,G}.

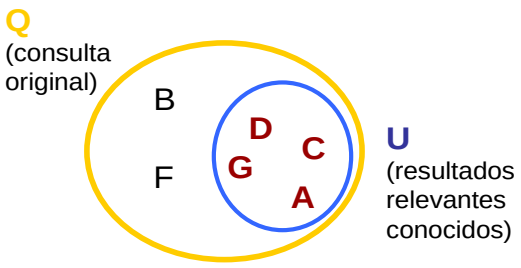


Figura 36: Ejemplo de estimación de cobertura y novedad. Conjunto de resultados relevantes conocidos U.

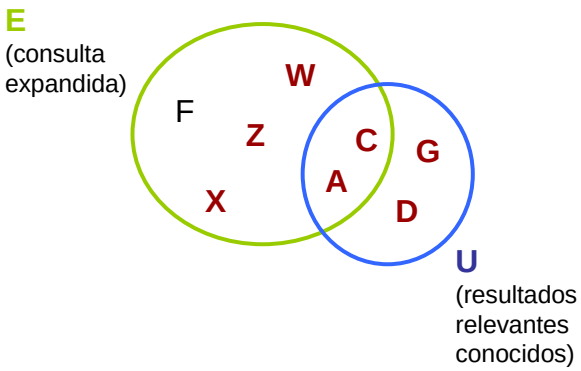


Figura 37: Ejemplo de estimación de cobertura y novedad. Conjunto de resultados relevantes desconocidos.

La consulta expandida (E) recupera los siguientes objetos E={X,A,C,Z,F,W}, cuyas evaluaciones de relevancia son {R,R,R,PR,N,PR}. Entonces como se representa en la Figura 37, el conjunto de objetos relevantes conocidos recuperados por la consulta expandida queda definido por E∩U=Rk={A,C} con |Rk|=2.

Luego, el conjunto de los objetos relevantes desconocidos Ru, contiene los objetos relevantes que no pertenecen a U, esto es Ru={X,Z,W} con |Ru|=3. Finalmente el resultado para las métricas de cobertura y novedad en el ejemplo es:

cobertura=2/4=0,5 (Ecuación 16).

novedad=3/5=0,60 (Ecuación 17).

Un alto nivel de cobertura indica que la expansión es capaz de recuperar la mayoría de los OA relevantes conocidos, es decir la mayoría de los OA relevantes obtenidos con la consulta original (sin expansión). Si el valor de cobertura=1 significa que la consulta expandida recuperó los mismos OA relevantes que la consulta sin expansión. Considerando que la lista de resultados de la expansión es la **unión** entre los resultados de la consulta original y las expansiones, si la cobertura es 1 indica que ambas listas contienen los mismos elementos. Un valor de cobertura=0 significa que la consulta expandida no recuperó ninguno de los objetos relevantes que fueron obtenidos de la consulta sin expansión (relevantes conocidos), es decir que Rk es vacío (Rk=∅) o bien que el conjunto U es vacío[18] (U=∅) lo que implica a su vez que Rk=∅, es decir la consulta original no recuperó ningún OA relevante.

Un alto nivel de novedad indica que la consulta expandida recupera mayor cantidad de OA relevantes nuevos, distintos a los recuperados sin la expansión. El valor novedad=1 significa que todos los OA relevantes que recuperó la consulta expandida son nuevos,

[18] Matemáticamente la fórmula de cobertura se indetermina por que el divisor es cero, es decir el conjunto U.

distintos a los recuperados sin la expansión, es decir $Rk=\emptyset$ y $Ru\neq\emptyset$. El valor de $novedad=0$ significa que la consulta expandida no recuperó OA relevantes distintos a los ya recuperados de la consulta sin expansión, $Ru=\emptyset$ o bien que la consulta expandida no recuperó resultados relevantes[19], $(Rk\cup Ru)=\emptyset$.

5.9. Limitaciones

Antes de discutir los resultados obtenidos de la expansión es necesario destacar las limitaciones de este caso de estudio:

- La definición de las consultas de prueba se realiza pensando que éstas deben representar las necesidades de un profesor en el diseño de un curso en el área de genética. Razón por lo cual, en nuestro caso de estudio se considera que una fuente válida para extraer dichas necesidades de información son las listas de contenidos tratados en los cursos de genética. La colección de syllabus fue restringida a programas publicados en la Web que son impartidos por instituciones de educación superior en el año académico 2009.

- La base de conocimiento para realizar la expansión es una ontología formal de dominio en genética.

  La ontología elegida en nuestro caso de estudio es GENE. Dicha ontología representa el conocimiento en el dominio extraído, validado y actualizado continuamente por la comunidad científica de esta área.

  Aún cuando es indiscutible la validez de dicha base de conocimiento, es preciso mencionar que la completitud del conocimiento representado en la ontología influye en los resultados de la expansión.

- El repositorio sobre el cual se prueba la expansión.

  MERLOT es un repositorio maduro y reconocido en comunidad e-learning. En particular, es uno de los pocos repositorios con una colección de OA en el área de genética que nos permitiría probar la expansión de consultas en dicha área.

  La relevancia de los resultados de la expansión esta directamente relacionada con los mecanismos de etiquetado, indexación y ranking implementados en el repositorio.

- La evaluación de la relevancia de los OA fue realizada por 3 expertos con conocimiento en genética y también en docencia.

[19] Matemáticamente la fórmula de novedad se indetermina por que el divisor es cero, es decir la unión de los resultados relevantes conocidos y desconocidos.

5.10. Resultados

A partir de las evaluaciones de relevancia de los OA recuperados, se calculan las métricas de cobertura, novedad, DCG y precisión para las *listas de resultados por tipo de expansión*. Cabe mencionar que, tal como fue explicado en la sección 4.1.7, dichas listas son generadas a partir de la unión entre las *listas individuales por tipo de expansión* y los resultados de la consulta original. A continuación, desde la Tabla 31 hasta la Tabla 42 se incluyen los resultados obtenidos para las consultas evaluadas.

Tabla 31: Resultados consulta Original, T3 *DNA replication*.

ranking	OA	Relevancia	Σ relevancia	precisión j-doc	DCG
1	85277	1	1	1	1
2	75865	1	2	1	2
3	271499	1	3	1	2,631
4	272460	1	4	1	3,131
5	79615	0	4	0,8	3,131
6	86024	1	5	0,833	3,518
7	80196	1	6	0,857	3,874
8	91364	1	7	0,875	4,207
9	299509	1	8	0,889	4,522
10	86519	1	9	0,9	4,823
precisión $\bar{x}$	0,9154				
DCG	4,823				

Tabla 32: Lista de resultados por tipo de expansión en la consulta T3 *DNA replication*. Expansión *is_a*: hermanos (isa_her) y *is_a*: hijos (isa_hij).

ranking	OA	Relevancia	∑ relevancia	precisión j-doc	DCG
1	85277	1	1	1	1
2	85278	0,5	1,5	0,75	1,5
3	75865	1	2,5	0,833	2,131
4	85280	0	2,5	0,625	2,131
5	271499	1	3,5	0,7	2,562
6	85285	1	4,5	0,75	2,949
7	272460	1	5,5	0,786	3,305
8	88124	0	5,5	0,688	3,305
9	79615	0	5,5	0,611	3,305
10	88706	0	5,5	0,55	3,305
Precisión $\bar{x}$	0,7293				
DCG	3,305				
Novelty	0,3333333				
Coverage	0,4444444				

OA	Relevancia	∑ relevancia	precisión j-doc	DCG
85277	1	1	1	1
90081	1	2	1	2
75865	1	3	1	2,631
85278	0,5	3,5	0,875	2,881
271499	1	4,5	0,9	3,312
79890	0	4,5	0,75	3,312
272460	1	5,5	0,786	3,668
85281	0,5	6	0,75	3,835
79615	0	6	0,667	3,835
89967	1	7	0,7	4,136
Precisión $\bar{x}$	0,8428			
DCG	4,136			
Novelty	0,5			
Coverage	0,444444			

Tabla 33: Resultados Consulta Original, T6 *chromosome*.

ranking	OA	relevancia	∑ relevancia	precisión j-doc	DCG
1	336262	1	1	1	1
2	325461	1	2	1	2
3	335495	1	3	1	2,631
4	335489	1	4	1	3,131
5	335478	1	5	1	3,562
6	90095	0,5	5,5	0,917	3,755
7	84206	1	6,5	0,929	4,111
8	75665	1	7,5	0,938	4,444
9	352312	1	8,5	0,944	4,759
10	86663	0,5	9	0,9	4,91
precisión $\bar{x}$	0,9628				
DCG	4,91				

Tabla 34: Lista de resultados por tipo de expansión en la consulta T6 *chromosome*. Expansión *is_a*: hermanos (isa_her) y *is_a*: padre (isa_exa).

ranking	OA	relevancia	∑ relevancia	precisión j-doc	DCG
1	336262	1	1	1	1
2	78960	0	1	0,5	1
3	325461	1	2	0,667	1,631
4	91401	0	2	0,5	1,631
5	335495	1	3	0,6	2,062
6	88403	0	3	0,5	2,062
7	335489	1	4	0,571	2,418
8	91398	0,5	4,5	0,563	2,585
9	335478	1	5,5	0,611	2,9
10	335450	1	6,5	0,65	3,201
precisión $\bar{x}$	0,6162				
DCG	3,201				
Novelty	0,2857143				
Coverage	0,5				

OA	relevancia	∑ relevancia	precisión j-doc	DCG
336262	1	1	1	1
89218	0,5	1,5	0,75	1,5
325461	1	2,5	0,833	2,131
85943	0,5	3	0,75	2,381
335495	1	4	0,8	2,812
335489	1	5	0,833	3,199
335478	1	6	0,857	3,555
90095	0,5	6,5	0,813	3,722
84206	1	7,5	0,833	4,037
75665	1	8,5	0,85	4,338
precisión $\bar{x}$	0,8319			
DCG	4,338			
Novelty	0,2			
Coverage	0,8			

Tabla 35: Resultados consulta original T15 *meiosis* Lista de resultados por tipo de expansión en la consulta T15 *meiosis*. Expansión *is_a*: hermanos (isa_her).

ranking	OA	Relevancia	□ relevancia	precisión j-doc	DCG
1	91401	1	1	1	1
2	84051	1	2	1	2
3	90943	1	3	1	2,631
4	75251	1	4	1	3,131
5	81501	1	5	1	3,562
6	91398	1	6	1	3,949
7	77749	1	7	1	4,305
8	86024	1	8	1	4,638
9	80042	1	9	1	4,953
10	82020	1	10	1	5,254
Precisión $\bar{x}$	1				
DCG	5,254				

OA	Relevancia	□ relevancia	precisión j-doc	DCG
91401	1	1	1	1
81127	0	1	0,5	1
84051	1	2	0,667	1,631
90943	1	3	0,75	2,131
78511	0	3	0,6	2,131
75251	1	4	0,667	2,518
81501	1	5	0,714	2,874
91398	1	6	0,75	3,207
243069	0,5	6,5	0,722	3,365
77749	1	7,5	0,75	3,666
Precisión $\bar{x}$	0,712			
DCG	3,666			
Novelty	0,125			
Coverage	0,7			

Tabla 36: Resultados Consulta Original, T16 *transcription*.

ranking	OA	relevancia	∑ relevancia	precisión j-doc	DCG
1	85278	1	1	1	1
2	85281	1	2	1	2
3	83048	0	2	0,667	2
4	82021	1	3	0,75	2,5
5	79615	0	3	0,6	2,5
6	215700	0,5	3,5	0,583	2,693
7	87817	0	3,5	0,5	2,693
8	82020	1	4,5	0,563	3,026
9	335489	1	5,5	0,611	3,341
10	335478	1	6,5	0,65	3,642
precisión $\bar{x}$	0,6924				
DCG	3,642				

Tabla 37: Lista de resultados por tipo de expansión en la consulta T16 *transcription*. Expansión *is_a*: hermanos (*isa_her*).

ranking	OA	relevancia	∑ relevancia	precisión j-doc	DCG
1	85278	1	1	1	1
2	85277	0,5	1,5	0,75	1,5
3	85281	1	2,5	0,833	2,131
4	85280	1	3,5	0,875	2,631
5	83048	0	3,5	0,7	2,631
6	88706	0	3,5	0,583	2,631
7	82021	1	4,5	0,643	2,987
8	224790	0	4,5	0,563	2,987
9	79615	0	4,5	0,5	2,987
10	75865	0,5	5	0,5	3,138
precisión $\bar{x}$	0,6947				
DCG	3,138				
Novelty	0,5				
Coverage	0,4285714				

Tabla 38: Lista de resultados por tipo de expansión en la consulta T16 *transcription*. Expansión *part_of*: el todo (*par_exa*) y *part_of*: las partes (*par_otr*).

ranking	OA	relevancia	∑ relevancia	precisión j-doc	DCG
1	85278	1	1	1	1
2	89622	0,5	1,5	0,75	1,5
3	85281	1	2,5	0,833	2,131
4	86015	1	3,5	0,875	2,631
5	83048	0	3,5	0,7	2,631
6	85217	1	4,5	0,75	3,018
7	82021	1	5,5	0,786	3,374
8	86663	0,5	6	0,75	3,541
9	79615	0	6	0,667	3,541
10	79775	1	7	0,7	3,842
precisión $\bar{x}$	0,7811				
DCG	3,842				
Novelty	0,625				
Coverage	0,428571				

OA	relevancia	∑ relevancia	precisión j-doc	DCG
85278	1	1	1	1
85346	0,5	1,5	0,75	1,5
85281	1	2,5	0,833	2,131
88706	0	2,5	0,625	2,131
83048	0	2,5	0,5	2,131
82021	1	3,5	0,583	2,518
224790	0	3,5	0,5	2,518
79615	0	3,5	0,438	2,518
238025	1	4,5	0,5	2,833
215700	0,5	5	0,5	2,984
precisión $\bar{x}$	0,6229			
DCG	2,984			
Novelty	0,333333			
Coverage	0,571429			

Tabla 39: Resultados Consulta Original, T17 *translation*

ranking	OA	relevancia	∑ relevancia	precisión j-doc	DCG
1	88706	0	0	0	0
2	224790	0	0	0	0
3	85281	1	1	0,333	0,631
4	88610	0	1	0,25	0,631
5	88006	0	1	0,2	0,631
6	88372	0	1	0,167	0,631
7	83577	0,5	1,5	0,214	0,809
8	82465	0	1,5	0,188	0,809
9	91428	0	1,5	0,167	0,809
10	88197	0	1,5	0,15	0,809
precisión $\bar{x}$	0,1669				
DCG	0,809				

Tabla 40: Lista de resultados por tipo de expansión en la consulta T17 *translation*. Expansión sinónimo exacto (*syn_exa*).

ranking	OA	relevancia	∑ relevancia	precisión j-doc	DCG
1	88706	0	0	0	0
2	83252	1	1	0,5	1
3	224790	0	1	0,333	1
4	85278	1	2	0,5	1,5
5	85281	1	3	0,6	1,931
6	85279	1	4	0,667	2,318
7	88610	0	4	0,571	2,318
8	86288	0,5	4,5	0,563	2,485
9	88006	0	4,5	0,5	2,485
10	79210	0,5	5	0,5	2,636
precisión $\bar{x}$	0,4734				
DCG	2,485				
Novelty	0,8333333				
Coverage	0,5				

Tabla 41: Lista de resultados por tipo de expansión en la consulta T17 *translation*. Expansión *is_a*: hermanos (*isa_her*) y *part_of*: el todo (*par_exa*).

ranking	OA	relevancia	∑ relevancia	precisión j-doc	DCG
1	88706	0	0	0	0
2	85277	1	1	0,5	1
3	224790	0	1	0,333	1
4	85278	1	2	0,5	1,5
5	85281	1	3	0,6	1,931
6	75865	0,5	3,5	0,583	2,124
7	88610	0	3,5	0,5	2,124
8	88006	0	3,5	0,438	2,124
9	271499	0,5	4	0,444	2,282
10	88372	0	4	0,4	2,282
precisión $\bar{x}$	0,4298				
DCG	2,124				
Novelty	0,8				
Coverage	0,5				

OA	relevancia	∑ relevancia	precisión j-doc	DCG
88706	0	0	0	0
89622	0,5	0,5	0,25	0,5
224790	0	0,5	0,167	0,5
86015	1	1,5	0,375	1
85281	1	2,5	0,5	1,431
85217	1	3,5	0,583	1,818
88610	0	3,5	0,5	1,818
86663	0	3,5	0,438	1,818
88006	0	3,5	0,389	1,818
79775	1	4,5	0,45	2,119
precisión $\bar{x}$	0,3652			
DCG	1,818			
Novelty	0,800			
Coverage	0,5			

Tabla 42: Lista de resultados por tipo de expansión en la consulta T17 *translation*. Expansión *part_of*: las partes (*par_otr*) y regulado por (*reg_por*).

ranking	OA	relevancia	∑ relevancia	precisión j-doc	DCG
1	88706	0	0	0	0
2	85278	1	1	0,5	1
3	224790	0	1	0,333	1
4	85346	0,5	1,5	0,375	1,25
5	85281	1	2,5	0,5	1,681
6	88610	0	2,5	0,417	1,681
7	238025	1	3,5	0,5	2,037
8	88006	0	3,5	0,438	2,037
9	83048	0	3,5	0,389	2,037
10	88372	0	3,5	0,35	2,037
precisión $\bar{x}$	0,3802				
DCG	2,037				
Novelty	0,750				
Coverage	0,5				

OA	relevancia	∑ relevancia	precisión j-doc	DCG
88706	0	0	0	0
86821	1	1	0,5	1
224790	0	1	0,333	1
85281	1	2	0,5	1,5
88610	0	2	0,4	1,5
88006	0	2	0,333	1,5
88372	0	2	0,286	1,5
83577	0,5	2,5	0,313	1,667
82465	0	2,5	0,278	1,667
91428	0	2,5	0,25	1,667
precisión $\bar{x}$	0,3193			
DCG	1,667			
Novelty	0,333			
Coverage	1,0			

En la Tabla 43 se resumen los resultados obtenidos de las expansiones en cada consulta. Las celdas vacías indican que una consulta no posee un tipo de expansión, esta situación se da porque no existen las relaciones en la ontología para ejecutar el tipo de expansión o porque no se recuperaron resultados en el repositorio.

Tabla 43: Resumen de los resultados de las métricas de novedad, cobertura y precisión para cada consulta y tipo de expansión.

		T15 meiosis	T3 DNA replication	T6 chromosome	T16 transcription	T17 translation	Varianza	Promedio
isa_her	Novedad	0,125	0,33333333	0,28571429	0,5	0,8	0,06568566	0,4088
	Cobertura	0,7	0,44444444	0,5	0,42857143	0,5	0,01178055	0,5146
	Precisión	0,712	0,7293	0,6162	0,6947	0,4298	0,01520906	0,6364
isa_hij	Novedad	*	0,5	*	*	*		0,5
	Cobertura	*	0,44444444	*	*	*		0,4444
	Precisión	*	0,8428	*	*	*		0,8428
isa_exa	Novedad	*	*	0,2	*	*		0,2
	Cobertura	*	*	0,8	*	*		0,8
	Precisión	*	*	0,8319	*	*		0,8319
par_exa	Novedad	*	*	*	0,625	0,8	0,0153125	0,7125
	Cobertura	*	*	*	0,42857143	0,5	0,00255102	0,4643
	Precisión	*	*	*	0,7811	0,3652	0,08648641	0,5732
par_otr	Novedad	*	*	*	0,33333333	0,75	0,08680556	0,5417
	Cobertura	*	*	*	0,57142857	0,5	0,00255102	0,5357
	Precisión	*	*	*	0,6229	0,3802	0,02945165	0,5016
syn_exa	Novedad	*	*	*	*	0,83333333		0,8333
	Cobertura	*	*	*	*	0,5		0,5
	Precisión	*	*	*	*	0,4734		0,4734

* indica que la consulta no tiene aplicado un tipo de expansión o bien, que no se recuperan OA desde el repositorio para dicha consulta.

5.11. Discusión de los resultados por tipo de expansión

En promedio las consultas expandidas obtienen un valor de novedad de 0.53 y de cobertura de 0.54. Estos resultados se interpretan de forma positiva puesto que los OA relevantes recuperados desde la expansión coinciden con alrededor de la mitad (0,54) de los OA relevantes recuperados de las consultas sin expansión. Al respecto se debe recordar que los resultados de las consultas expandidas están unidos a los resultados de las consultas originales. Además un 53% de los OA relevantes recuperados de las expansiones son nuevos, es decir distintos a los recuperados de las consultas sin expansión.

Los mejores resultados de novedad, cobertura y precisión se obtuvieron con la consulta T17: *translation*, en la Tabla 44 se listan los recursos recuperados por la consulta original y algunos de los tipos de expansiones aplicados. En la tabla se destacan los 2 OA evaluados como relevantes para la consulta original. Es de resaltar que en las listas de los resultados de las expansiones sólo se incluyen los resultados de la consulta de los términos expandidos, antes de unirlos a los resultados de la consulta original (*lista individual por tipo de expansión*). Como se presenta en la Figura 38, de los diez recursos recuperados con la consulta original un 80% de ellos se refieren a la *traducción de un lenguaje a otro*, en cambio más del 90% de los recursos recuperados de las consultas expandidas se refieren a la *traducción como parte del proceso de expresión genética*. Esto se explica porque el término buscado en esta consulta posee diferentes interpretaciones dependiendo del contexto donde se utilice, por ejemplo lingüístico o genético. Este caso es bastante frecuente debido a las características polisémicas de algunas palabras.

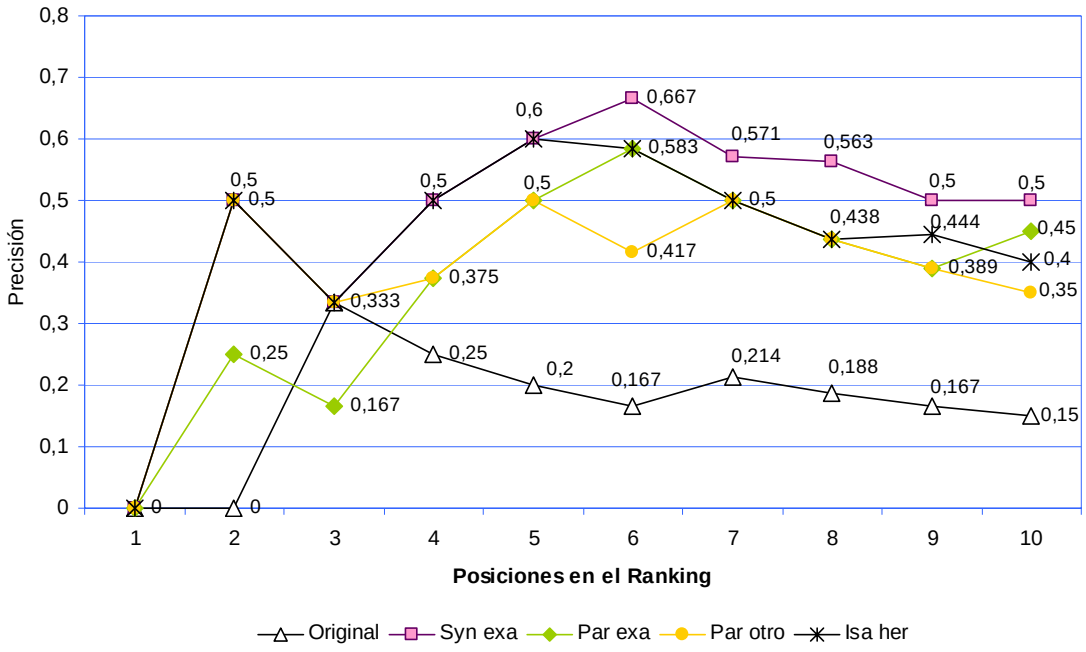
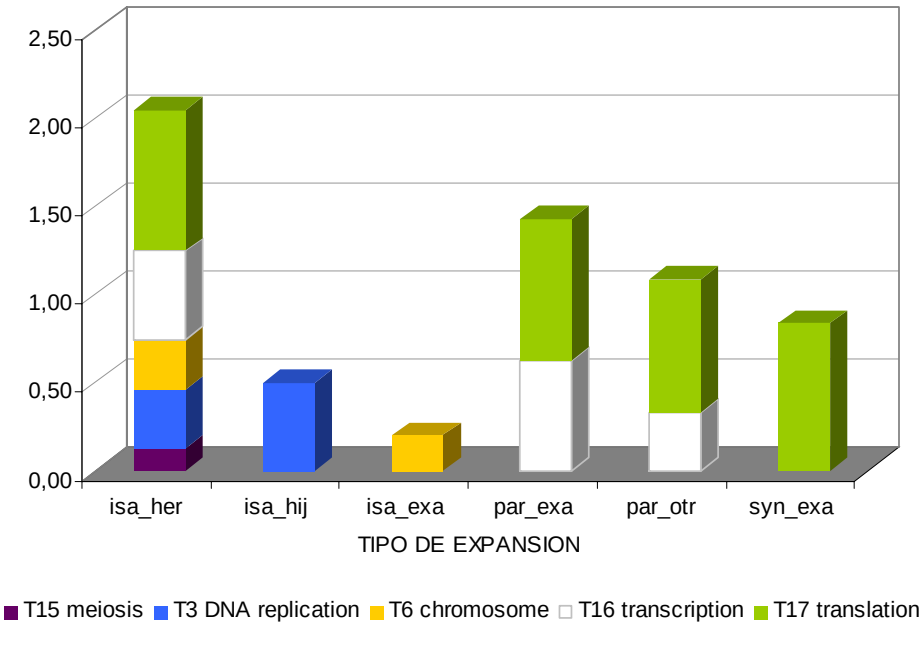


Figura 38: Niveles de precisión de los resultados recuperados por la consulta T17, original y expansiones.

Tabla 44: Lista de OA recuperados por la consulta original T17: *translation* y las listas individuales para algunos tipos de expansiones.

ranking	Consulta original	Lista individual por tipo de expansión		
		Expansión *isa_her*	Expansión *par_exa*	Expansión *syn_exa*
1	Ojala que llueva cafe	DNA Replication (Semiconservative Replication)	Biotechnology Workshop	Protein Translation
2	eStroke Animated Chinese Characters	Transcription (DNA to RNA)	The Virtual Library of Biochemistry, Molecular Biology and Cell Biology	Transcription (DNA to RNA)
3	**RNA Processing (post-transcriptional modifications)**	Biology Tutorials for Metabolism and Genetics	Kimball's Biology Pages	The Genetic Code
4	Chinese Love Poetry and Folklore	RNA Processing (post-transcriptional modifications)	Genome Bioinformatics	Blazing a Genetic Trail
5	SignWritingSite	DNA Replication Video	UCLA Molecular Biology Tutorials	Molecular Biochemistry
6	Chinese Character Flashcards	Max Hunter Folk Song Collection	StarBiogene	Biology Tutorials for Metabolism and Genetics
7	**Connecting Concepts: Biotechnology**	Lew-Port DNA Replication	Potent Biology: Stem Cells, Cloning, and Regeneration	Biology Textbook (On-Line)
8	German Vocabulary Trainer	Virtual Cell Animations	Genes can be turned on and off: regulation of gene expression	Protein Synthesis Dance
9	MendelWeb	Biology Flash Animations		Lew-Port's Biology Place--Animated Reviews
10	Speak Mandarin in Five Hundred Words	Flash animations in science		Protein Synthesis

Debido al tamaño del experimento y a la cantidad de expansiones que recuperan OA en el repositorio, no es posible obtener datos concluyentes respecto a si existen diferencias entre los resultados de los distintos tipos de expansión. Algunos tipos de expansión tienen resultados en 1 de las 5 consultas evaluadas. Los tipos de expansión regulado por (*reg_por*), otros sinónimos (*syn_all*) y regula a (*reg_a*) no tienen evaluación ya que no fueron aplicados en las consultas evaluadas o bien, no recuperan OA del repositorio.



La Figura 39 sintetiza los resultados de la métrica de novedad para cada consulta y tipo de expansión. Como se puede observar, el tipo de expansión más veces aplicado es *is_a*: hermanos. Por el contrario, el tipo de expansión sinónimo exacto *syn_exa* sólo fue aplicado en la consulta T17.

La métrica de novedad es siempre mayor que cero, lo cual indica que las consultas expandidas, independiente del tipo, contribuyen en la recuperación de OA relevantes distintos a los resultados recuperados de la consulta sin expansión. Dicho de otra forma las consultas expandidas permiten acceder a OA relevantes a los que sin la expansión no sería posible acceder.

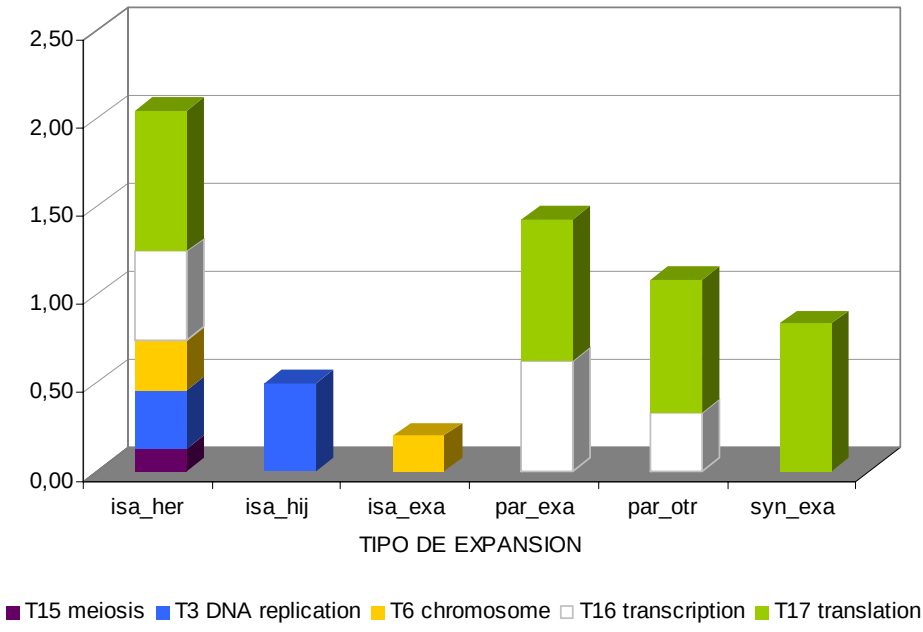


Figura 39: Novedad para cada consulta y tipo de expansión.

Los mejores niveles de novedad se obtienen en los tipos de expansión *part_of*: el todo, *part_of*: las partes e sinónimo *syn_exa*. Los dos primeros se encuentran aplicados en las consultas T16 y T17, si analizamos sólo estas dos consultas el tipo de expansión *is_a*: hermano tiene un índice de novedad mayor al tipo *part_of*: las partes.

La Figura 40 presenta los resultados de la métrica de cobertura. Como se observa, la cobertura es siempre mayor a 0. Ello se explica dado que la lista de los resultados de las consultas expandidas es la unión entre los resultados de las consultas de los términos expandidos y de la consulta original. Cabe recordar que en promedio la cobertura es 0,54 lo que indica que la repetición entre los resultados de estas listas es baja.

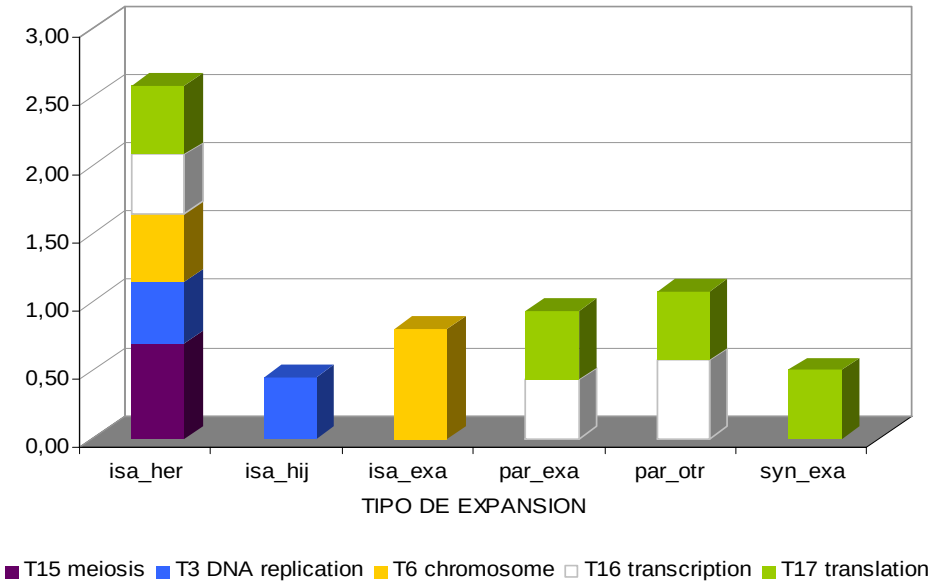


Figura 40: Cobertura para cada consulta y tipo de expansión.

Los niveles más altos de cobertura se dan en los tipos de expansión *is_a*: hermano y *is_a*: padre, es decir que en estos casos existe mayor coincidencia entre los OA relevantes recuperados con y sin expansión.

En términos de la precisión, los niveles más altos se encuentran en los tipos de expansión *is_a*: hermanos y *is_a*: padre, en estos casos se superan los niveles de precisión obtenidos con la consulta original (sin expansión). En general las expansiones presentan niveles de precisión iguales o inferiores a los obtenidos con las consultas sin expansión (ver Figura 41).

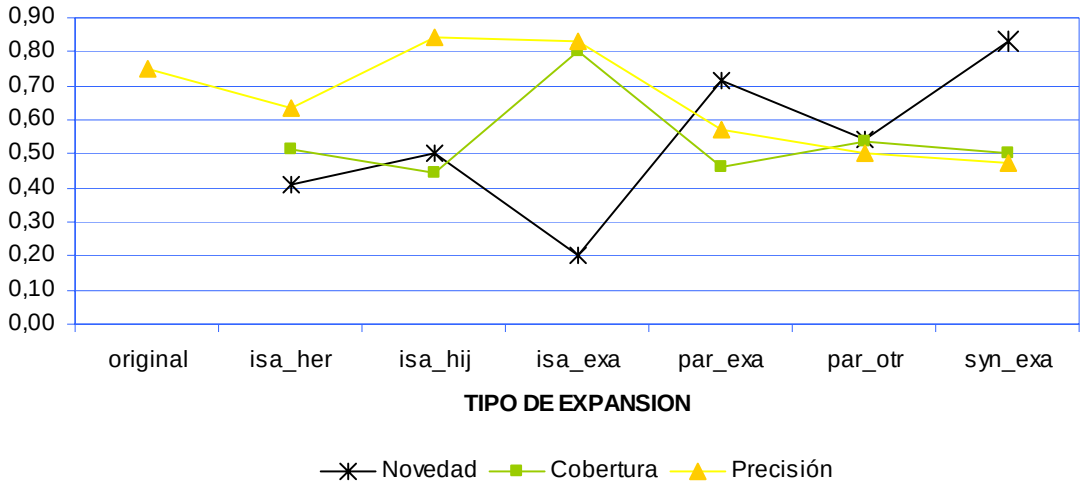


Figura 41: Promedio de cobertura, novedad y precisión.

5.12. Discusión de los resultados independientemente del tipo de expansión

En el análisis anterior se comentaron las diferencias entre los resultados obtenidos de los distintos de expansión, a continuación analizamos los resultados independientemente del tipo de expansión. Tal como fue descrito en la sección 4.1.7, se procesan 3 tipos de listas:

- *Lista de resultados individuales por tipo de expansión*. Se incluyen los resultados de la consulta de los términos expandidos.
- *Lista de resultados por tipo de expansión*. Es la intercalación de los resultados individuales de cada tipo de expansión y los resultados de la consulta original.
- *Lista única de expansión*. Es la intercalación de los resultados individuales de todos los tipos de expansión y los resultados de la consulta original.

Como se observa en la Tabla 45, en cualquiera de las consultas e independiente del tipo de expansión del que se trate las *listas individuales por tipo de expansión* tienen mejores niveles de novedad que cualquiera de las otras listas de resultados. Esta situación es completamente explicable puesto que en las otras dos listas al intercalar los resultados de las expansiones con los resultados de la consulta original se aumenta el número de resultados relevantes conocidos (`|Rk|`) y se reduce o mantiene el número de resultados relevantes no conocidos (`|Ru|`). El aumento de `|Rk|` se debe a la incorporación de los OA obtenidos de la consulta original y la reducción o mantención de `|Ru|` se debe a la disminución de los OA obtenidos de la expansión.

Por ejemplo en la consulta T16 la *lista individual del tipo de expansión is_a*: hermanos (`isa_her`) tiene un índice de novedad de `0,83` (`|Rk|=1` y `|Ru|=5`), en cambio al unir esta

lista con los resultados de la consulta original en la lista `original +isa_her` (*lista por tipo de expansión)* el índice de novedad baja a `0,2` (`|Rk|=3` y `|Ru|=3`). En promedio la novedad baja `0,5` puntos.

Por otro lado, en cualquiera de las consultas e independiente del tipo de expansión del que se trate las *listas por tipo de expansión* tienen menores niveles de novedad que los obtenidos en las *listas únicas de expansión*. Siguiendo el ejemplo de la consulta T16 cualquiera de las *listas únicas de expansión* tienen un nivel de novedad de `0,75` (`|Rk|=2` y `|Ru|=6`) en cambio en el mejor caso la *lista por tipo de expansión part_of*: el todo alcanza un nivel de novedad de `0,63` (`|Rk|=3` y `|Ru|=5`). Lo anterior se traduce en que las listas que reúnen los resultados de todas las expansiones ofrecen mayor cantidad de resultados relevantes distintos a los obtenidos de la consulta original (sin expansión).

Tabla 45: Novedad y Cobertura en las distintas listas de resultados.

Consulta	Tipo Lista[d]	Lista	Rk[a]	Ru[b]	U[c]	Cobertura	Novedad
T15	*Lista individuales por tipo de expansión*	isa_her	2	3	4	0,50	0,60
	Lista por tipo de expansión	original+isa_her	7	1	10	0,70	0,13
T16	*Listas individuales por tipo de expansión*	isa_her	1	5	7	0,14	0,83
		par_exa	0	5	7	0,00	1,00
		par_otr	1	1	7	0,14	0,50
	Listas por tipo de expansión	original+isa_her	3	3	7	0,43	0,50
		original+par_exa	3	5	7	0,43	0,63
		original+par_otr	4	2	7	0,57	0,33
	listas única de expansión	original+isa_her+par_exa+par_otr	2	6	7	0,29	0,75
		original+par_exa+par_otr+isa_her	2	6	7	0,29	0,75
		original+par_otr+par_exa+isa_her	2	6	7	0,29	0,75
T17	*Listas individuales por tipo de expansión*	isa_her	1	7	2	0,50	0,88
		par_exa	0	6	2	0,00	1,00
		par_otr	1	6	2	0,50	0,86
		syn_exa	0	7	2	0,00	1,00
	Listas por tipo de expansión	original+isa_her	1	4	2	0,50	0,80
		original+par_exa	1	4	2	0,50	0,80
		original+par_otr	1	3	2	0,50	0,75
		original+syn_exa	1	5	2	0,50	0,83
	Listas única de expansión	original+isa_her+par_exa+par_otr+syn_exa	1	7	2	0,50	0,88
		original+par_exa+par_otr+syn_exa+isa_her	1	7	2	0,50	0,88
		original+par_otr+par_exa+isa_her+sy	1	7	2	0,50	0,88

Consulta	Tipo Lista[d]	Lista	Rk[a]	Ru[b]	U[c]	Cobertura	Novedad
		n_exa					
		original+syn_exa+ par_exa+par_otr+i sa_her	1	7	2	0,50	0,88
T3	*Listas individuales por tipo de expansión*	isa_her	0	6	9	0,00	1,00
		isa_hij	1	4	9	0,11	0,80
	Listas por tipo de expansión	original+isa_her	4	2	9	0,44	0,33
		original+isa_hij	4	4	9	0,44	0,50
	Listas única de expansión	original+isa_her+is ahijo	4	4	9	0,44	0,50
		original+isahijo+isa _her	4	4	9	0,44	0,50
T6	*Listas individuales por tipo de expansión*	isa_exa	0	2	10	0,00	1,00
		isa_her	0	4	10	0,00	1,00
	Listas por tipo de expansión	original+isa_exa	8	2	10	0,80	0,20
		original+isa_her	5	2	10	0,50	0,29
	listas única de expansión	original+isa_exa+is a_her	4	4	10	0,40	0,50
		original+isa_her+is a_exa	4	4	10	0,40	0,50

a: Promedio de resultados relevantes conocidos obtenidos en la consulta expandida, más detalle en la sección 5.8.
b: Promedio de resultados relevantes NO conocidos obtenidos en la consulta expandida
c: Promedio de resultados relevantes conocidos obtenidos en la consulta original (sin expansión)
d: Tipo de listas de resultados: individuales por tipo de expansión, por expansión o únicas.

original: lista de resultados de la consulta sin expansión
isa_exa: Expansión de los conceptos padres del concepto buscado.
isa_her: Expansión de los conceptos hermanos del concepto buscado.
isa_hij: Expansión de los conceptos hijos del concepto buscado.
par_exa: Expansión de los conceptos que contienen al concepto buscado.
par_otr: Expansión de los conceptos que están contenidos en el mismo concepto que el concepto buscado.
syn_exa: Expansión de los conceptos sinónimos exactos.

La Figura 42 sintetiza los resultados presentados anteriormente. Las *listas únicas de expansión* obtienen un índice de novedad que supera o iguala al obtenido en las *listas por tipo de expansión*. Por ejemplo la lista que une los resultados de las listas `original +isa_her +par_exa +par_otr` tiene un nivel de novedad igual a `0,88`, este índice es superior al obtenido por cualquiera de las listas para las expansiones *is_a*: hermanos (`original +isa_her`, `novedad=0,41`), *part_of*: el todo (`original +par_exa`, `novedad=0,715`), *part_of*: las partes (`original +par_otr`, `novedad=0,54`), *sinónimo*: exacto (`original +syn_exa`, `novedad=0,83`).

La mejora en los niveles de novedad se explica porque la intercalación toma los resultados mejor posicionados en el ranking de cada una de las listas a intercalar.

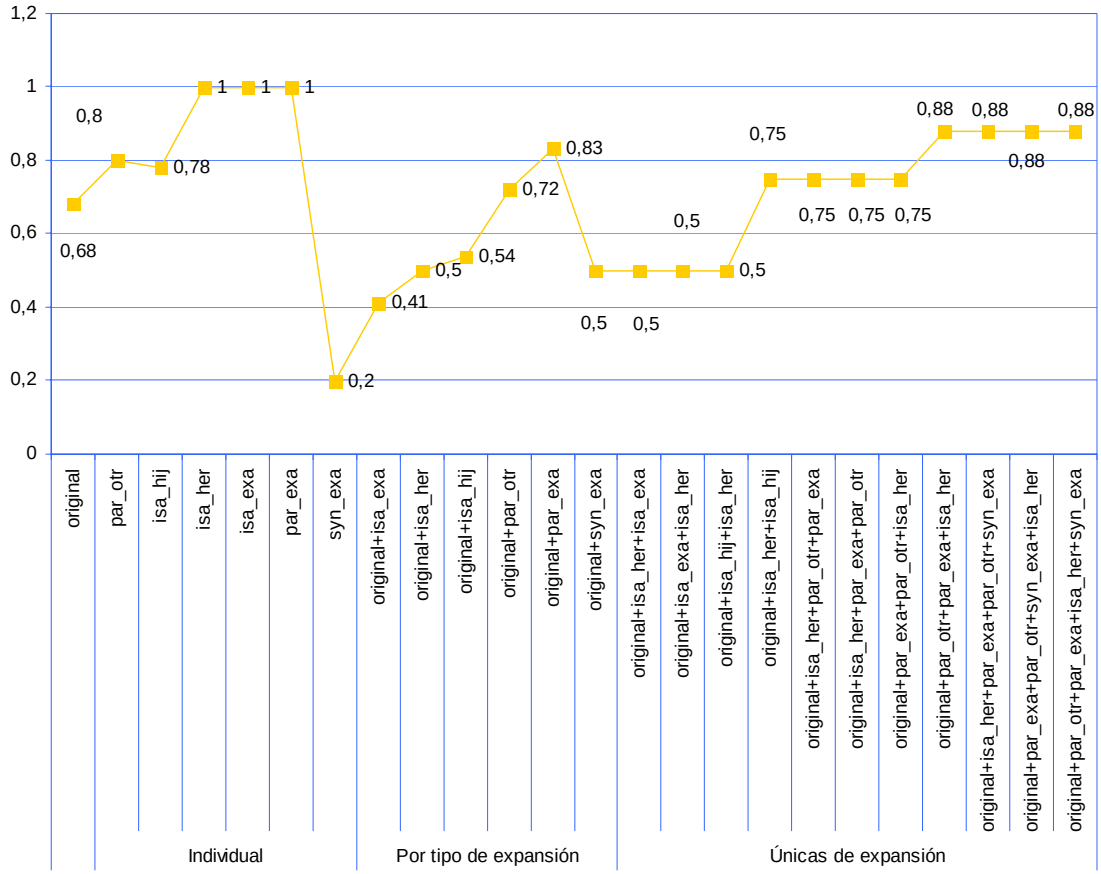


Figura 42: Resumen de la métrica de Novedad para las distintas listas de análisis.

El orden en la intercalación equivale a asignar pesos distintos por cada tipo de expansión, e intercalar en función de las posiciones de ranking ponderadas. Por ejemplo, la consulta T16 *transcription* obtuvo resultados para 3 tipos de expansión *is_a*: hermanos (`isa_her`), *part_of*: el todo (`par_exa`) y *part_of*: las partes (`par_otr`). En la unión de los resultados de cada tipo de expansión con la lista de resultados de la consulta original se puede privilegiar la expansión `isa_her` por sobre los demás tipos de expansión, y a su vez ponderar que la expansión `par_otr` es mejor a la expansión `par_exa`, en este caso los resultados serán unidos en el siguiente orden:

```
original +isa_her +par_otr +par_exa
```

Al respecto, cabe destacar que las mejoras en los resultados de novedad obtenidos en las *listas únicas de expansión* son independientes al orden de intercalación de cada tipo de expansión. Siguiendo el ejemplo anterior, los resultados de novedad de las siguientes seis posibles listas (`n!=3!=6`, donde `n:` es la cantidad de tipos de expansión aplicados a la consulta) no varían y continúan siendo superiores a los obtenidos en las *listas por tipo de expansión*.

```
original +isa her +par exa +par otr
original +isa_her +par_otr +par_exa
original +par otr +par exa +isa her
original +par_otr +isa_her +par_exa
```

```
original +par exa +par otr +isa her
original +par_exa +par_otr +isa_her
```

Por otra parte, asumiendo que el algoritmo de ranking del repositorio ubica en las primeras posiciones los resultados de mayor relevancia, podemos suponer que *la novedad será mayor en la medida que se intercalen mayor cantidad de tipos de expansión.*

Como se ejemplifica en la Figura 42, existen consultas con resultados para 2, 3 y 4 tipos de expansión, el nivel de novedad en cada uno aumenta en la medida que aumentan la cantidad de tipos de expansión intercalados.

En la Figura 43, a diferencia del comportamiento observado en el nivel de novedad de las listas de resultados, en la medida que aumentan la cantidad de tipos de expansión aplicadas en una consulta la precisión no presenta una variación importante.

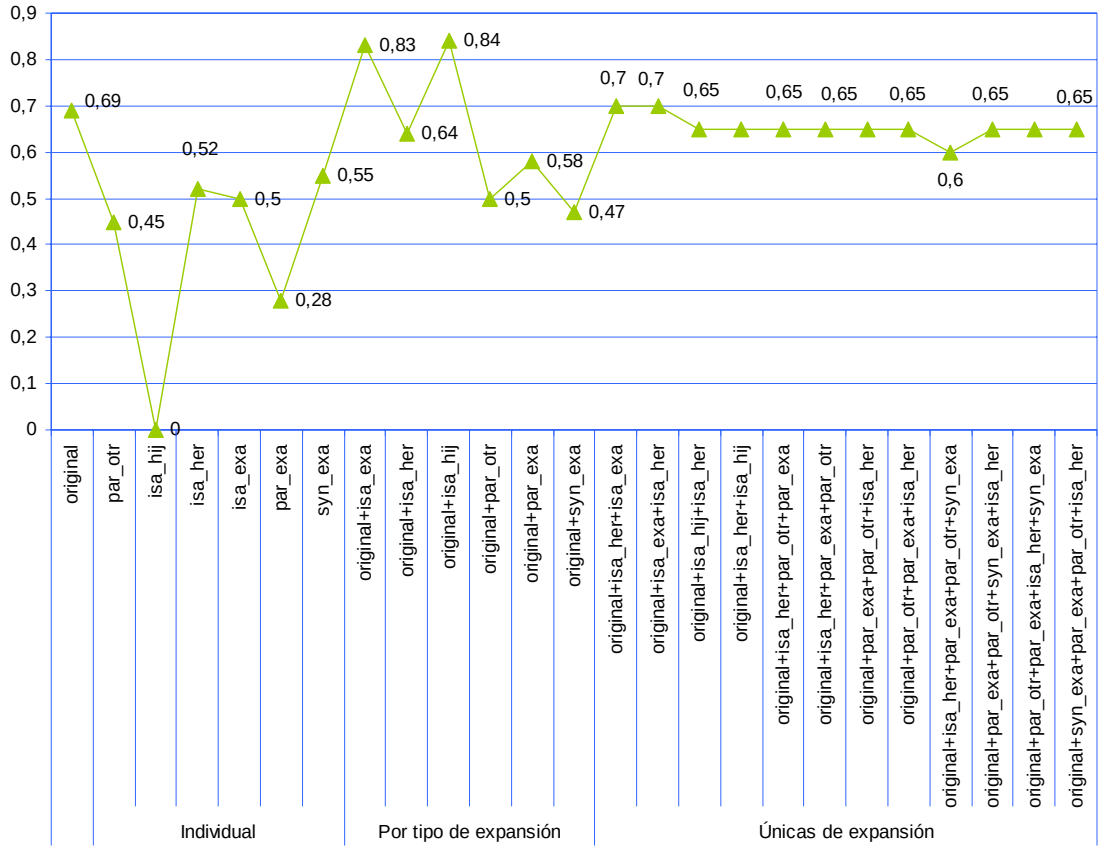


Figura 43: Resumen de la métrica de Precisión para las distintas listas de análisis.

En este capítulo se ha evaluado la estrategia propuesta para la expansión de consultas basada en ontología aplicada en el contexto de la búsqueda de OA en repositorios. Es de destacar que en el ámbito de la recuperación de OA en repositorio no existe una colección de prueba estándar que nos permita contrastar los resultados de esta propuesta, por lo tanto para la evaluación de la estrategia propuesta se diseñó un experimento dentro del dominio de genética. La evaluación de la relevancia de los resultados de la expansión fue realizada por 3 expertos en el dominio y el análisis de la

concordancia y asociación entre sus evaluaciones es realizado por medio del análisis de *Kappa de Cohen* y el coeficiente de correlación de *Spearman*. Finalmente, la efectividad de la propuesta de expansión se evalúa a partir de las métricas de cobertura y novedad aplicadas a los resultados recuperados de las consultas con y sin expansión.

El trabajo realizado permite examinar la consecución de los objetivos planteados en el tercer capítulo, actividad que se abordaría en el capítulo siguiente.

Capítulo 6.Conclusiones

En este capítulo se presentan las conclusiones finales de esta tesis, para ello se verifica que tanto la hipótesis como los objetivos derivados han sido alcanzados a lo largo de la investigación. De igual forma se describen las aportaciones y las líneas de trabajo futuras derivadas directamente de esta tesis. Por último, se presentan las publicaciones logradas durante el periodo de formación doctoral tanto en eventos científicos, capítulos de libros y revistas indexadas.

6.1. Verificación de la Hipótesis y los Objetivos

La hipótesis planteada en esta tesis doctoral es:

> *El uso de ontologías para la expansión de consultas por parte de tutores/diseñadores instruccionales en repositorios de objetos de aprendizaje puede mejorar la novedad de los resultados de las consultas.*

Dada la hipótesis planteada el objetivo principal definido para ser alcanzado en la tesis es:

> *Proponer una estrategia para la expansión de consultas basada en ontologías de dominio que permita al diseñador instruccional obtener resultados relevantes desde los repositorios de objetos de aprendizaje.*

A continuación, para cada uno de los objetivos específicos definidos para alcanzar el objetivo principal se describen los elementos de esta memoria a partir de los cuales es posible verificarlos.

> *(O1) Revisar las técnicas de expansión de consultas, específicamente aquellas basadas en ontologías, y dentro de estas, las utilizadas o propuestas para la búsqueda en repositorios de objetos de aprendizaje.*

En el capítulo 2 se documenta el estudio y análisis exhaustivo de las propuestas existentes para la expansión de consultas en general y específicamente, aquellas basadas en modelos de conocimiento. Dentro de este mismo capítulo también se documentan los elementos teóricos relacionados con las áreas de interés de esta tesis. Entre ellos están los estándares de metadatos, los repositorios de objetos de aprendizaje, la recuperación de información y los modelos de representación de conocimiento.

La revisión sistemática y la documentación obtenida de ella proporcionan los fundamentos necesarios y suficientes para avanzar con éxito en el desarrollo de esta investigación.

(O2) Establecer los tipos de relaciones de la ontología y los tipos de expansión aplicables en el ámbito de la recuperación de objetos de aprendizaje en repositorios.

De acuerdo al contexto de la investigación definido en el capítulo 3, en el capítulo 4 quedan especificadas las bases de la propuesta para la expansión de consultas basada en ontologías. Estas bases se refieren a los tipos de relaciones representados en la ontología que son utilizados en la expansión, la distancia semántica y los tipos de expansiones, que de acuerdo a los dos aspectos mencionados antes, definen los conceptos que serán agregados a la consulta del diseñador instruccional para recuperar OA relevantes.

(O3) Especificar un procedimiento para resolver el problema de correspondencia entre el concepto buscado y los conceptos modelados en la ontología, dentro del contexto de la búsqueda de objetos de aprendizaje.

Uno de los problemas que se generan implícitamente del uso de modelos de conocimientos en la expansión automática de consultas es la falta de correspondencia exacta entre la consulta y los conceptos modelados en la ontología. En el capítulo 4 se describe el procedimiento a partir del cual pretendemos dar solución a este problema dentro del contexto de la investigación.

En síntesis, cuando no existe correspondencia exacta entre el concepto consultado y algún concepto de la ontología, se seleccionan en la ontología los conceptos que contienen al concepto buscado, esto sin considerar los *stopwords*. A partir de dichos conceptos se selecciona el concepto más cercano a la consulta y que agregue menor ambigüedad. Es decir, se selecciona el concepto que contiene al concepto buscado y que a la vez incluye una menor cantidad de palabras nuevas a las contenidas en el concepto consultado.

(O4) Especificar una solución a los problemas o limitaciones detectadas para la implementación de la expansión de consultas basadas en ontologías en los repositorios de objetos de aprendizaje.

Después de analizar las funcionalidades y mecanismos de búsqueda disponibles en algunos de los repositorios de OA más utilizados, han sido detectadas algunas limitaciones para implementar la expansión de consultas basada en ontologías. Estas

limitaciones guardan relación con el lenguaje de consultas utilizado, cantidad máxima de términos de la cadena de búsqueda y el uso de conectores lógicos.

De acuerdo a lo anterior, en el capítulo 4 se especifica una forma que nos permite superar estas limitaciones y abstraernos del repositorio que se utilice para aplicar la estrategia de expansión de consultas.

(O5) Plantear, diseñar e implementar la estrategia de expansión de consultas basadas en ontologías en el ámbito de la recuperación de objetos de aprendizaje en repositorios.

En el capítulo 4, tras especificar los distintos elementos que definen y especifican la expansión de consultas (O2, O3 y O4), se plantea la estrategia que los reúne y permite su aplicación en la búsqueda de objetos de aprendizaje en repositorios. Además, en el capítulo 4 se describe la arquitectura técnica y funcional de la propuesta, y la funcionalidad del prototipo implementado. Dicho prototipo permite el pre-procesamiento de la cadena de búsqueda, la expansión de la consulta según cada uno de los tipos definidos, la búsqueda en el repositorio y el procesamiento de los resultados recuperados por éste.

Nuestra propuesta es independiente del sistema de recuperación del repositorio (indexación, almacenamiento, ranking y formato de consultas) ya que las consultas se expresan como una búsqueda conjuntiva por cada concepto y los resultados se procesan en el formato XML compatible con el esquema de metadatos IEEE-LOM.

(O6) Comprobar la efectividad de la propuesta, su marco de aplicabilidad, sus limitaciones y sus posibilidades de extensión hacia otros dominios.

En el capitulo 5 se detalla el experimento diseñado para evaluar la propuesta de expansión de consultas basada en ontologías aplicado a la búsqueda en repositorios de OA. Las consultas de prueba se obtienen a partir de los conceptos incluidos en los listados de contenidos de los programas académicos de cursos en un dominio, en concreto las consulta de prueba quedan definidas por los conceptos que poseen mayor frecuencia dentro de la colección de syllabus.

En nuestro experimento se utilizaron 24 syllabus de cursos en el área de genética publicados en la Web por instituciones de educación superior para el periodo 2009. La relevancia de los resultados es evaluada por expertos del dominio y el nivel de acuerdo de su percepción es validado estadísticamente a través del análisis Kappa (κ) y el coeficiente de correlación de Spearman (δ).

Los resultados de la expansión se evalúan a partir de las métricas de novedad y cobertura aplicadas tanto a las listas de OA obtenidos individualmente por cada tipo de expansión como a las listas de OA obtenidos de la unión de los distintos tipos de expansión.

6.2. Conclusiones

Tal como ha sido analizado por (Ali & Khan, 2008; Bhogal et al., 2007; Billerbeck & Zobel, 2004; Kang & Kim, 2003) los métodos de expansión mejoran el desempeño del sistema de RI dependiendo del tipo de consulta al que se aplique. Incluso el método de expansión más apropiado puede depender también del tipo de consulta al que es aplicado. En nuestro estudio el tipo de consulta se encuentra caracterizada por los siguientes aspectos:

- La consulta se enmarca dentro de un dominio de conocimiento específico en el cual se encuentra el curso e-learning que el profesor diseña.
- La consulta es menos ambigua puesto que el usuario es un profesor con conocimiento en el dominio en el cual se encuentra el curso e-learning que diseña.
- La intención del usuario, implícita en la consulta, se enmarca dentro del contexto del e-learning y específicamente, asociada a la tarea de diseñar o componer un nuevo recurso de aprendizaje o un curso e-learning.
- El resultado esperado de la consulta es un conjunto de recursos digitales creados con un propósito educativo, es decir son objetos de aprendizaje y no cualquier otro tipo de material o información disponible en la Web.

Basado en los resultados obtenidos del experimento podemos decir que:

Cuando un profesor busca recursos de aprendizaje en un repositorio, para diseñar un curso e-learning o componer un nuevo recurso, la expansión de consultas basada en ontología le permite acceder a recursos relevantes que sin la expansión no podrían ser recuperados.

Los resultados obtenidos de nuestro caso de estudio no permiten concluir respecto a las diferencias entre cada uno de los tipos de expansión. Los tipos de expansión no son siempre aplicables a todas las consultas, existen casos en que el concepto buscado en la ontología no posee las relaciones necesarias para la expansión, incluso es frecuente que aunque un concepto posea expansiones éstas no recuperen resultados en el repositorio que permitan su análisis.

Por otro lado, es interesante destacar que cuando se unen los resultados en una única lista con todos los tipos de expansión, independientemente del orden en el que sean intercalados, los niveles de novedad y precisión mejoran. Lo anterior nos permite suponer que *si el algoritmo de ranking del repositorio ubica en las primeras posiciones los resultados de mayor relevancia para la consulta entonces la novedad será mayor en la medida que se apliquen más tipos de expansión, es decir en la medida que se intercalen los resultados obtenidos de más tipos de expansión.*

La replicación de este caso de estudio depende de los OA recuperados por el repositorio. Se entiende que estos resultados pueden variar debido al crecimiento del contenido del repositorio, no obstante según hemos constatado, los resultados entregados por el repositorio también varían en función del mecanismo de acceso

utilizado para la consulta. En nuestro experimento el repositorio MERLOT fue consultado utilizando un servicio RESTFul, pero al ejecutar las mismas pruebas utilizando la herramienta SQITest o directamente en el portal Web del repositorio, los resultados para una misma consulta no fueron exactamente iguales.

Los resultados de la expansión de consultas *se ven afectados por la calidad de la ontología, el sistema de recuperación implementado en el repositorio y las características de la colección de OA almacenados en éste.* Si el conocimiento modelado en la ontología es incompleto o inconsistente, los términos extraídos del modelo también lo serán y en consecuencia, los recursos relevantes obtenidos con la expansión podrían disminuir. De forma similar, si los recursos almacenados en el repositorio son de baja calidad, se encuentran mal etiquetados o mal indexados, los recursos recuperados de la expansión podrían ser menos relevantes para el usuario.

La extensión de los resultados obtenidos en nuestra investigación a otras áreas de conocimiento requiere la existencia de una ontología formal y validada, y además requiere que un repositorio posea una colección de OA dentro del nuevo dominio y provea los mecanismos para acceder a ellos.

Desde el punto de vista técnico, el tamaño de las ontologías de dominio impone un desafío para la implementación de un servicio web que de soporte a la estrategia de expansión de consultas propuesta. Según nuestra experiencia el módulo de expansión implementado a través de los frameworks y APIs actualmente disponibles para el manejo de ontologías (jena y sparql), requiere gran cantidad de recursos computacionales aunque en nuestro caso sólo implique la consulta del modelo.

6.3. Aportaciones

Las principales contribuciones de nuestra investigación son:

- Una estrategia para la expansión de consultas basadas en ontologías definida en el contexto específico de la recuperación de objetos de aprendizaje en repositorios. Como fue ampliamente discutido en los apartados anteriores, la mayor parte de las investigaciones y propuestas previas se aplican a la recuperación de información en la Web. La delimitación del contexto de esta tesis ha permitido caracterizar de forma específica nuestra investigación en función del tipo de usuario, nivel de conocimientos de éste en el dominio de la consulta, intención de la consulta y tipo de recurso digital buscado.
- Una arquitectura técnica y funcional independiente de la ontología utilizada como base de conocimiento para extraer los conceptos expandidos. La expansión de consultas se realiza utilizando los tipos de relaciones comunes en cualquier modelo ontológico, además se incluyen relaciones propias del dominio. Estas últimas también pueden ser configurables puesto que varían de un dominio otro o de una ontología a otra.

- Una arquitectura técnica y funcional independiente del repositorio de objetos de aprendizaje donde se ejecute la búsqueda. La cadena de búsqueda de la consulta expandida es tratada de forma tal que pueda ser procesada por cualquier repositorio, es decir, las consultas se expresan como una búsqueda simple conjuntiva por cada concepto (original y expandido). Por otro lado, los resultados entregados por el repositorio se procesan en el formato XML compatible con el esquema estándar de metadatos IEEE-LOM. Dicho estándar, o un perfil de aplicación basado en él, es utilizado en la mayor parte de los repositorios y protocolos de intercambio de OA.
- Un proceso sistemático para la evaluación de los efectos de la expansión de consultas en el contexto de la búsqueda de OA en repositorios. Dado que no existe una colección de prueba estándar definida para la recuperación de OA en repositorios, en esta investigación se ha definido un proceso de evaluación que permite contrastar los resultados obtenidos por la estrategia de expansión propuesta. El conjunto de consultas de prueba se obtiene a partir de los conceptos incluidos en los listados de contenidos de los syllabus de cursos en un dominio. Las consultas de prueba quedan definidas por los conceptos que poseen mayor frecuencia en la colección de programas académicos. En nuestro experimento se utilizaron 24 syllabus de cursos en el área de genética publicados en la Web por instituciones de educación superior en el periodo 2009. Dentro de nuestro proceso de evaluación, la relevancia de los resultados es evaluada por expertos, donde la concordancia entre sus percepciones es validada estadísticamente.

6.4. Líneas de Trabajo Futuras

A partir de esta investigación se abre la posibilidad de llevar a cabo un buen número de proyectos en las siguientes líneas de trabajo:

- Analizar si existen diferencias significativas entre los distintos tipos de expansiones (*is_a* versus *part-of*) o tipos de relaciones (relaciones léxicas y ontológicas) que se utilicen en la expansión. La ejecución de nuevos experimentos, específicos según los tipos de relaciones o expansiones, en el mismo domino de conocimiento permitirá complementar los hallazgos realizados en la presente investigación.
- Mejorar el mecanismo definido para resolver el problema de la falta de correspondencia entre el concepto consultado y los conceptos modelados en la ontología.
- Evaluar la efectividad de las distintas estrategias definidas para establecer la correspondencia entre el concepto consultado y los conceptos modelados en la ontología cuando no existe una correspondencia exacta. La efectividad de las estrategias, evaluada por expertos del dominio, puede ser medida como el grado de cercanía conceptual entre el concepto consultado y el concepto seleccionado como equivalente. En la medida que el concepto seleccionado como equivalente es más cercano al concepto buscado suponemos que los resultados de la expansión serán

similares a los obtenidos cuando existe una correspondencia exacta entre la consulta y los conceptos de la ontología.

- Analizar la validez de los resultados obtenidos en nuestro caso de estudio cuando se replica en un dominio de conocimiento distinto. Manteniendo las constantes de nuestro caso de estudio y variando sólo el dominio de conocimiento en el cual se enmarcan las consultas de prueba es posible analizar si los resultados en cuanto a la novedad, cobertura y precisión presentan un comportamiento similar.
- Integrar la estrategia de expansión de consultas propuesta con técnicas para la expansión según criterios pedagógicos relacionados con la audiencia y el contexto donde serán usados los OA. Nuestra propuesta de expansión de consultas se enfoca hacia la temática o el tópico de la consulta, no obstante un OA puede ser relevante según la temática pero no ser apropiado al nivel de profundidad del curso, los conocimientos previos de los alumnos, la interactividad o participación de los alumnos con los recursos, entre otros criterios. Suponemos incluso que algunos de estos criterios pueden tener incidencia en la estrategia de expansión propuesta, por ejemplo en el tipo de relación ontológica o expansión más adecuada para extraer los conceptos expandidos. En la Figura 44 se representa la ampliación de la arquitectura propuesta según esta línea de trabajo.

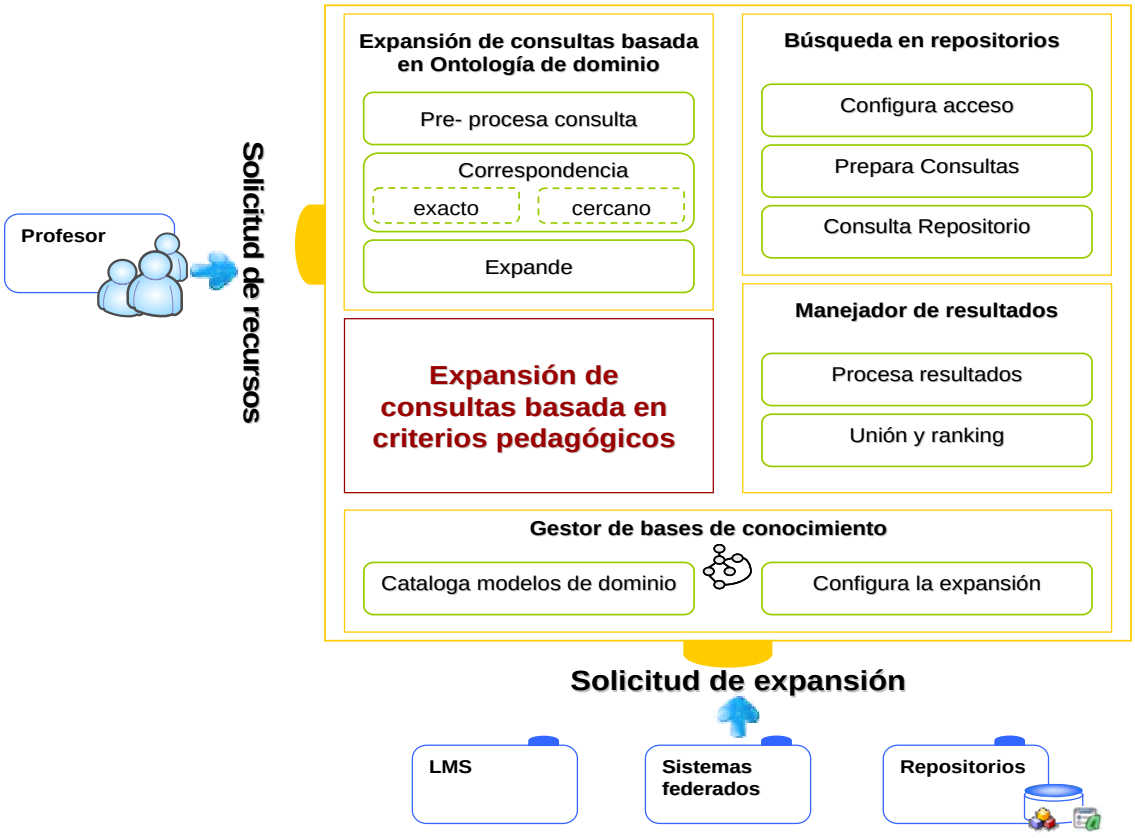


Figura 44: Arquitectura propuesta ampliada según ésta línea de trabajo futura

6.5. Publicaciones y Contribuciones derivadas

Entre los materiales de divulgación científica generados durante nuestra investigación, destacamos en primer lugar la publicación en la revista *Knowledge based sytems*. En este artículo se presenta nuestra estrategia de expansión de consultas basada en ontologías de dominio y los resultados obtenidos de la evaluación.

Tipo : Revista Internacional

Autores : Alejandra A. Segura, Salvador-Sánchez, Elena García-Barriocanal, Manuel Prieto

Título : **An empirical analysis of ontology-based query expansion for learning resource searches using MERLOT and the Gene Ontology**

Revista : Knowledge based Systems, 24 (2011), 119-133.

Tipo : Revista Internacional

Autores : Segura, A.; Vidal, C.; Campos, P.; Menéndez, V.; Prieto, M.

Título : **Exploring Characterizations of Learning Object Repositories Using Data Mining Techniques**

Revista : The Electronic Library, (29) issue 2, 2011 **– In Press**

6.6. Otras Publicaciones

A continuación se detallan otras publicaciones logradas durante el periodo de formación doctoral tanto en eventos científicos, capítulos de libros y revistas indexadas. Estas publicaciones se dirigen hacia las distintas temáticas relacionadas con esta tesis, resumiendo el trabajo desde la exploración en el área de investigación hasta la concreción del contexto específico de investigación. Algunos de las temáticas tratadas son: calidad de los recursos de aprendizaje, calidad de los metadatos de los OA almacenados en repositorios, búsqueda, meta búsqueda y recomendación de OA y expansión de consultas basadas en ontologías.

6.6.1. Revistas o Capítulos de libros

Título : **Knowledge-based architecture for instructional engineering**

Autores : Christian L. Vidal, Alejandra A. Segura, Víctor H. Menéndez, Manuel E. Prieto

Revista : International Journal of Knowledge and Learning 2009 - Vol. 5, No.3/4 pp. 371 - 388

Título : **A Recommender System architecture for instructional engineering.**

Autores : Manuel Prieto, Víctor Menéndez, Alejandra Segura, Christian Vidal

Libro : Lecture Notes in Computer Science, Volume 5288/2008, pp. 314-321. Springer Berlin / Heidelberg (2008) ISBN 978-3-540-87780-6, 2008

Título : **Characterizations of Learning Object Repositories Using Data Mining Techniques.**

Autores : Segura, A., Vidal, C., Menéndez, V., Zapata, A., Prieto, M.

Libro : Metadata and Semantic Research. F. Sartori, M.Á. Sicilia, and N. Manouselis (Eds.): MTSR 2009, CCIS 46. Springer Berlin / Heidelberg, pp. 215-225 (2009). ISBN 978-3-642-04589-9, 2009.

6.6.2. Congresos internacionales

Título : **Query Expansion based on Domain Ontology for Learning Objects Search**

Autores : Alejandra Segura N., Christian Vidal and Manuel Prieto,

Congreso : 3rd IEEE International Conference on Computer Science and Information Technology (IEEE ICCSIT 2010)

Lugar : Chengdu, China

Fecha : 9 - 11 Julio 2010

Título : **Quality in Learning Objects: Evaluating compliance with metadata Standards.**

Autores : Christian Vidal C., Alejandra Segura N., Pedro Campos S., Salvador Sánchez-Alonso.

Congreso : MTSR 2010 - Fourth International Conference on Metadata and Semantics Research,

Lugar : Alcalá de Henares.

Fecha : 20-22 Octubre 2010.

Título : **Characterizing metadata in Learning Object Repositories.**

Autores : Alejandra Segura N., Christian Vidal C., Victor Menéndez, Alfredo Zapata, Manuel Prieto.

Congreso : MTSR 2009 - Third International Conference on Metadata and Semantics Research,

Lugar : Milán, Italia.

Fecha : 30 Septiembre – 02 Octubre 2009

Título : **A Recommender System Architecture for Instructional Engineering**

Autores : Manuel Prieto, Victor Menéndez, Alejandra Segura y Christian Vidal

Congreso : World Summit on the Knowledge Society

Lugar : Atenas, Grecia

Fecha : 24-28 Septiembre 2008

Título : **Búsqueda y Composición de Objetos de Aprendizaje.**

Autores : Alejandra A. Segura, Víctor Menéndez, Manuel E. Prieto.

Congreso : Proccedings of X International Symposium on Computers in Education SIIE 2008. Universidad de Salamanca

Lugar : Salamanca, España

Fecha : 01-03 Octubre 2008

Título : **Evaluación de la Calidad del Software para el Aprendizaje**

Autores : Alejandra A. Segura; Christian L. Vidal; Manuel E. Prieto

Congreso : Proccedings of X International Symposium on Computers in Education SIIE 2008. Universidad de Salamanca

Lugar : Salamanca, España

Fecha : 01-03 Octubre 2008

Título : **Metadata and Ontologies in Learning Resources Design**

Autores : Christian Vidal C., Alejandra Segura N., Víctor Menéndez D., Alfredo Zapata G.2 Manuel Prieto M.

Congreso : World Summit on the Knowledge Society

Lugar : Corfu, Grecia

Fecha : 22-24 Septiembre 2010

Título : **An Approach to Metadata Generation for Learning Objects**

Autores : Victor Menendez D., Alfredo Zapata G., Christian Vidal C., Alejandra Segura N. , Manuel Prieto M.

Congreso : World Summit on the Knowledge Society

Lugar : Corfu, Grecia

Fecha : 22-24 Septiembre 2010

6.6.3. Congresos nacionales

Título : **Calidad en Objetos de Aprendizaje**

Autores : Christian L. Vidal; Alejandra A. Segura; Manuel E. Prieto

Congreso : V Simposio Pluridisciplinar sobre Diseño y Evaluación de Contenidos Educativos Reutilizables, SPEDECE 08.

Lugar : Salamanca, España

Fecha : 20-21 Octubre 2008

Referencias

Abbas, Z., Umer, M., Odeh, M., McClatchey, R., Ali, A., & Farooq, A. (2005). *A semantic grid-based e-learning framework (SELF).* Paper presented at the CCGRID '05: Proceedings of the Fifth IEEE International Symposium on Cluster Computing and the Grid (CCGrid'05) - Volume 1, 11-18.

Abdelali, A., Cowie, J., & Soliman, H. S. (2007). Improving query precision using semantic expansion. *Information Processing and Management: an International Journal, 43*(3), 705-716.

ADL. (2003). Sharable Courseware Object Reference Model (SCORM), Version 1.3, Advanced Distributed Learning. from http://www.scormsoft.com/scorm. Retrieved Septiembre 2010

AENOR. (2008). Perfil de Aplicación LOM (LOM-ES), V.1.0 G. G.-S. 36, Asociación Española de Normalización y Certificación. from http://www.educa.madrid.org/cms_tools/files/ac98a893-c209-497a-a4f1-93791fb0a643/lom-es_v1.pdf. Retrieved Agosto 2010

Agirre, E., & Rigau, G. (1996). *Word sense disambiguation using Conceptual Density.* Paper presented at the Proceedings of the 16th conference on Computational linguistics, 16-22.

Akritidis, L., Katsaros, D., & Bozanis, P. (2008). *Effective Ranking Fusion Methods for Personalized Metasearch Engines.* Paper presented at the PCI '08: Proceedings of the 2008 Panhellenic Conference on Informatics, 39-43.

Akritidis, L., Voutsakelis, G., Katsaros, D., Bozanis, P., & Greeee, T. (2007). QuadSearch: A Novel Metasearch Engine. *Current Trends in Informatics*, 453-466.

Al-Khalifa, H. S., & Davis, H. C. (2006). *The evolution of metadata from standards to semantics in E-learning applications.* Paper presented at the HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia, 69-72.

Ali, W., & Khan, S. (2008). *Ontology Driven Query Expansion in Data Integration.* Paper presented at the SKG '08: Proceedings of the 2008 Fourth International Conference on Semantics, Knowledge and Grid, 57-63.

Baader, F., Calvanese, D., McGuinness, D. L., Nardi, D., & Patel-Schneider, P. F. (2007). *The Description Logic Handbook: Theory, Implementation and Applications*: Cambridge University Press.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*: Addison-Wesley-Longman.

Bastiaan, M. K., Lalanne, A., & Shamseldin, S. (2003, August 5-8, 2003). *MERLOT Federated Search Technologies.* Paper presented at the Merlot International Conference, Vancouver, British Columbia, Canada.

Berners-Lee, T. (1998). Semantic Web Road Map [Electronic Version]. Retrieved Septiembre 2010 from http://www.w3.org/DesignIssues/Semantic.html.

Bhogal, J., Macfarlane, A., & Smith, P. (2007). A review of ontology based query expansion. *Information Processing and Management: an International Journal, 43*(4), 866-886.

Billerbeck, B., & Zobel, J. (2004). *Questioning query expansion: an examination of behaviour and parameters.* Paper presented at the In Proceedings of the fifteenth australasian database conference, ADC 2004, CRPIT, 69-76.

Blackman, N., & Koval, J. (2000). Interval estimation for Cohen's kappa as a measure of agreement. *Statistics in Medicine, 19*(5), 723-741.

Borlund, P. (2003). The concept of relevance in information retrieval. *Journal of the American Society for Information Science and Technology, 54*(10), 913-925.

Borst, W. N. (1997). *Construction of engineering ontologies for knowledge sharing and reuse.* Unpublished PhD, University of Twente, Netherlands.

Brin, S., & Page, L. (1998). *The anatomy of a large-scale hypertextual Web search engine.* Paper presented at the WWW7: Proceedings of the seventh international conference on World Wide Web 7, 107-117.

Budanitsky, A., & Hirst, G. (2001). *Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures.* Paper presented at the In Workshop on Wordnet and other lexical resources, Second meeting of the North American chapter of the association for computational linguistics, Pittsburgh

Calegari, S., & Pasi, G. (2008). *Personalized Ontology-Based Query Expansion.* Paper presented at the WI-IAT '08: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 256-259.

Cechinel, C., Sánchez-Alonso, S., Sicilia, M.-Á., & Sartori, F. (2009). Empirical Analysis of Errors on Human-Generated Learning Objects Metadata. In M.-Á. Sicilia & N. Manouselis (Eds.), *Metadata and Semantic Research* (pp. 60-70): Springer-Verlag Berlin Heidelberg 2009.

CEN. (2005). A Simple Query Interface Specification for Learning Repositories (SQI), CWA 15454:2005 E, Comité Européen de Normalisation. from ftp://ftp.cenorm.be/PUBLIC/CWAs/e-Europe/WS-LT/cwa15454-00-2005-Nov.pdf. Retrieved Septiembre 2010

Clark & Parsia. (2010). Pellet: OWL 2 Reasoner for Java (Version 2.2.1) [Ontology Reasoner]. on Line. Available http://clarkparsia.com/pellet/

Clark, J., & DeRose, S. (1999). XML Path Language -XPath. on Line: Consorcio World Wide Web (W3C). from http://www.w3.org/TR/xpath/. Retrieved Enero 2009.

Croft, W. B. (1986). *User-specified domain knowledge for document retrieval.* Paper presented at the SIGIR '86: Proceedings of the 9th annual international ACM SIGIR conference on Research and development in information retrieval, 201-206.

Chandrasekaran, B., Josephson, J. R., & Benjamins, V. R. (1999). What Are Ontologies, and Why Do We Need Them? *IEEE Intelligent Systems, 14*(1), 20-26.

Chli, M., & De Wilde, P. (2006). Internet search: Subdivision-based interactive query expansion and the soft semantic web. *Applied Soft Computing, 6*(4), 372-383.

Christensen, E., Curbera, F., Meredith, G., & Weerawarana, S. (2001). Web Services Description Language (WSDL), *Version 1.1*. on Line: Consorcio World Wide Web (W3C). from http://www.w3.org/TR/wsdl. Retrieved Septiembre 2010.

Chu, W. W., Liu, Z., & Mao, W. (2002). Textual Document Indexing and Retrieval via Knowledge Sources and Data Mining. *Communication of the Institute of Information and Computing Machinery (CIICM), 5*(2).

Dehors, S., & Zucker, C. F. (2006). *Reusing Learning Resources based on Semantic Web Technologies.* Paper presented at the ICALT '06: Proceedings of the Sixth IEEE International Conference on Advanced Learning Technologies, 859-863.

Demartini, G., & Mizzaro, S. (2006). A Classification of IR Effectiveness Metrics. In *Advances in Information Retrieval* (Vol. 3936/2006, pp. 488-491): Springer Berlin / Heidelberg.

Department of Computer Science at University of Manchester. (2004). Hoolet-OWL DL reasoner [Ontology Reasoner]. on Line. Available http://owl.man.ac.uk/hoolet/

Devedziz, V. (2006). Ontological Engineering for Semantic Web-Based Education. In *Semantic and education.* (Vol. 12, pp. 353): Springer US.

Díaz-Galiano, M. C., Martín-Valdivia, M. T., & Ureña-López, L. A. (2009). Query expansion with a medical ontology to improve a multimodal information retrieval system. *Comput. Biol. Med., 39*(4), 396-403.

Diehl, A. D., Lee, J. A., Scheuermann, R. H., & Blake, J. A. (2007). Ontology development for biological systems: immunology. *Bioinformatics, 23*(7), 913-915.

Ding, Y., & Foo, S. (2002a). Ontology Research and Development. Part 2 - A Review of Ontology Mapping and Evolving. *Journal of Information Science, 28*(5), 375-388.

Ding, Y., & Foo, S. (2002b). Ontology research and development. Part I: a review of ontology generation. *Journal of information science, 28*(2), 123-136.

DMCI. (2003). Dublin Core Metadata Element Set (DC), Versión 1.1, Dublin Core Metadata Initiative. from http://dublincore.org/documents/dces/. Retrieved Septiembre 2010

Dolog, P., Simon, B., Nejdl, W., & Klobucar, T. (2008). Personalizing access to learning networks. *ACM Trans. Internet Technol., 8*(2), 1-21.

Dunlop, M. D. (1997). The effect of accessing nonmatching documents on relevance feedback. *ACM Transactions on Information Systems (TOIS), 15*(2), 137-153.

ERIC. (2010). Education Resources Information Center. *Thesaurus* Retrieved Access ^, 2010, from http://www.eric.ed.gov/

Farance, F. (2003). IEEE LOM Standard Not Yet Ready For Prime Time. *IEEE Computer Society and Learning Technology Task Force (LTTF), 5*(1).

French, J. C., Powell, A. L., Gey, F., & Perelman, N. (2001). *Exploiting a controlled vocabulary to improve collection selection and retrieval effectiveness.* Paper presented at the Proceedings of the tenth international conference on Information and knowledge management, 199-206.

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*: Springer.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition, 5*(2), 199-220.

Guarino, N. (1997). Understanding, building and using ontologies. *Int. J. Hum.-Comput. Stud., 46*(2-3), 293-310.

Haarslev, V., & Möller, R. (2001). *RACER System Description.* Paper presented at the IJCAR '01: Proceedings of the First International Joint Conference on Automated Reasoning, 701-706.

Haarslev, V., Möller, R., & Turhan, A.-Y. (1998). *Implementing an ALCRP(D) ABox Reasoner - Progress Report.* Paper presented at the Proc. DL-98 International Description Logic Workshop 1998, June 6 - June 8, Trento, Italy, 82-86.

Haarslev, V., Möller, R., & Turhan, A. Y. (1999). *RACE User's Guide and Reference Manual Version 1.1.*

Hatala, M., Richards, G., Eap, T., & Willms, J. (2004a). *The eduSource Communication Language: implementing open network for learning repositories and services.* Paper presented at the Proceedings of the 2004 ACM symposium on Applied computing, 19-27.

Hatala, M., Richards, G., Eap, T., & Willms, J. (2004b). *The interoperability of learning object repositories and services: standards, implementations and lessons learned.* Paper presented at the Proceedings of the 13th international World Wide Web conference on Alternate track papers, 957-962.

Hernández, H., & Sáiz N, M. (2007). Ontologías mixtas para la representación conceptual de objetos de aprendizaje. *Procesamiento del lenguaje natural, 38*.

Hersh, W., Hickam, D., Haynes, R. B., & McKibbon, K. A. (1994). A performance and failure analysis of saphire with a medline test collection. *Journal of the American Medical Informatics Association, 1*, 51-60.

Hewlett-Packard Development Company. (2008). A Semantic Web Framework for Java - Jena (Version 2.5.6) [Java toolkit for developing semantic web applications]. on Line. Available http://jena.sourceforge.net/

Higgs, P., Meredith, S., & Hand, T. (2003). *Technology for sharing: Researching learning objects and digital rights management.*

Hilera, J. R., Bengochea, L., de Madariaga, R. S., de Mesa, J. A. G., & Martínez, J. J. (2005). *Aplicación de técnicas de Ingeniería Lingüística en sistemas de e-learning basados en objetos de aprendizaje.* Paper presented at the II Simposio Pluridisciplinar sobre Diseño y Evaluación de Contenidos Educativos Reutilizables.

Ho, J., & Tang, R. (2001). *Towards an optimal resolution to information overload: an infomediary approach.* Paper presented at the Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work, 91-96.

Huang, L. (2000). *A Survey On Web Information Retrieval Technologies.*

Huang, X., Huang, Y. R., & Wen, M. (2005). *A dual index model for contextual information retrieval.* Paper presented at the SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, 613-614.

Huang, Y.-F., & Hsu, C.-H. (2008). PubMed smarter: Query expansion with implicit words based on gene ontology. *Knowledge-Based Systems., 21*(8), 927-933.

Hull, D. A., & Grefenstette, G. (1996). *A detailed analysis of english stemming algorithms* (Technical report). Xerox. http://www.xrce.xerox.com/Research-Development/Publications/1996-023/%28language%29/eng-GB.

IMS. (2002). Learning Resource Meta-data (IMS-LRM), Specification V1.3, Global Learning Consortium lrm. from http://www.imsglobal.org/metadata/mdv1p3/imsmd_bestv1p3.html. Retrieved Septiembre 2010

IMS. (2003a). Digital Repositories Interoperability (IMS-DRI), Final Specification Version 1.0, Global Learning Consortium dri. from http://www.imsglobal.org/digitalrepositories/driv1p0/imsdri_infov1p0.html. Retrieved Septiembre 2010

IMS. (2003b). Learning Design (IMS-LD), Specification V1.0, Global Learning Consortium ld. from http://www.imsglobal.org/learningdesign/index.html. Retrieved Septiembre 2010

ISO. (1986). Guidelines for the establishment and development of monolingual thesauri (ISO 2788:1986), International Organization for Standarization. from http://www.iso.org/iso/catalogue_detail.htm?csnumber=7776 Retrieved Septiembre 2010

Jarvelin, K., & Kekalainen, J. (2000). *IR evaluation methods for retrieving highly relevant documents.* Paper presented at the SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, 41-48.

JDOM TM Project. (2009). JDOM (Version 1.1.1) [Java-based solution for accessing, manipulating, and outputting XML data from Java code]. on Line. Available http://www.jdom.org/

Jike, G., & Yuhui, Q. (2008). *Concept Similarity Matching Based on Semantic Distance.* Paper presented at the Proceedings of the 2008 Fourth International Conference on Semantics, Knowledge and Grid.

Joho, H., Sanderson, M., & Beaulieu, M. (2004). *A Study of User Interaction with a Concept-based Interactive Query Expansion Support Tool.* Paper presented at the Advances in Information Retrieval, 26th European Conference on Information Retrieval, 42-56.

Jovanovic, J., Knight, C., Gasevic, D., & Richards, G. (2006). *Learning Object Context on the Semantic Web.* Paper presented at the Proceedings of the Sixth IEEE International Conference on Advanced Learning Technologies.

Kaczmarek, J., & Landowska, A. (2006). Model of distributed learning objects repository for a heterogenic internet environment. *Interactive Learning Environments, 14*(1).

Kang, I.-H., & Kim, G. (2003). *Query type classification for web document retrieval.* Paper presented at the Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval.

Khare, R., & Taylor, R. N. (2004). *Extending the Representational State Transfer (REST) Architectural Style for Decentralized Systems.* Paper presented at the ICSE '04: Proceedings of the 26th International Conference on Software Engineering, 428-437.

Krovetz, R., & Um, C. S. (1993). *Viewing Morphology as an Inference Process.*

Kwasnik, B. H. (1999). The Role of Classification in Knowledge Representation and Discovery. *Library Trends, 48*(1), 22-47.

Lagoze, C., & Van de Sompel, H. (2002). The Open Archives Initiative Protocol for Metadata Harvesting, *Version 2.0*. on Line. from http://www.openarchives.org/OAI/openarchivesprotocol.html. Retrieved Septiembre 2010.

Lamarca Lapuente, M. J. (2006). *Hipertexto, el nuevo concepto de documento en la cultura de la imagen.* Universidad Complutense de Madrid.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Lee, M.-C., Tsai, K. H., & Wang, T. I. (2008). A practical ontology query expansion algorithm for semantic-aware learning objects retrieval. *Computers & Education, 50*(4), 1240-1257.

Lee, M.-C., Yen Ye, D., & Wang, T. I. (2005). *Java Learning Object Ontology.* Paper presented at the Proceedings of the Fifth IEEE International Conference on Advanced Learning Technologies, 538-542.

Lempel, R., & Moran, S. (2000). *The stochastic approach for link-structure analysis (SALSA) and the TKC effect.* Paper presented at the Proceedings of the 9th international World Wide Web conference on Computer networks : the international journal of computer and telecommunications netowrking, 387-401.

Lempel, R., & Moran, S. (2005). Rank-Stability and Rank-Similarity of Link-Based Web Ranking Algorithms in Authority-Connected Graphs. *Information Retrieval, 8*(2), 245-264.

Lovins, J. (1968). Development of a Stemming Algorithm. *Mechanical Translation and Computational Linguistics, 11*, 22-31.

LTSC. (2002). IEEE Learning Object Metadata (LOM), Draft Standard 1484.12.1, Learning Technology Standards Committee. from http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf. Retrieved Noviembre de 2007

Ma, L., Chen, L., Gao, Y., & Yang, Y. (2009). *Ontology Based Query Expansion in Vertical Search Engine.* Paper presented at the FSKD '09: Proceedings of the 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 285-289.

Mandala, R., Tokunaga, T., & Tanaka, H. (1999). *Combining Multiple Evidence from Different Types of Thesaurus for Query Expansion.* Paper presented at the SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*: W. H. Freeman Co.

McGreal, R. (2004). Learning objects: A practical definition. *International Journal of Instructional Technology and Distance Learning (IJITDL), 9*(1), 21-32.

McGreal, R., Adelsberger, H. H., Kinshuk, P., Pawlowski, J. M., & Sampson, D. G. (2008). A Typology of Learning Object Repositories. In *Handbook on Information Technologies for Education and Training* (pp. 5-28): Springer-Verlag Berlin Heidelberg 2009.

McGuinness, D. L., & van Harmelen, F. (2004). OWL Web Ontology Language, *Overview*. on Line: Consorcio World Wide Web (W3C). from http://www.w3.org/TR/owl-features/. Retrieved Agosto 2010.

McMartin, F. (2004). MERLOT: A Model for User Involvement in Digital Library Design and Implementation. *Journal of Digital Information, 5*(3).

Menéndez, V., Vidal, C., Urzaiz, J., & Prieto, M. (2008). *A Web Services-Based Model For The Composition Of Reusable Learning Objects.* Paper presented at the Inted2008, International Technology, Education And Development Conferenceiated, International Association Of Technology, Education And Development.

Meng, S., Brown, D. E., Ebbole, D. J., Torto-Alalibo, T., Yee Oh, Y., Deng, J., et al. (2009). Gene Ontology annotation of the rice blast fungus, Magnaporthe oryzae. *BMC Microbiology, 9*(Suppl 1:S8).

MERLOT. (2010). Multimedia educational resource for learning and online teaching. Retrieved Access ^, 2008, from http://www.merlot.org/merlot/index.htm

Mizzaro, S. (1996). *How many relevances in IR?* Paper presented at the In Proceedings of the Workshop Information Retrieval and Human Computer Interaction, 57-60.

Morales, E., Gil, A., & García, F. (2007). *Arquitectura para la Recuperación de Objetos de Aprendizaje de calidad en Repositorios Distribuidos.* Paper presented at the XII Jornadas de Ingeniería del Software y Bases de Datos.

Morgan, E. L. (2004). An Introduction to the Search/Retrieve URL Service (SRU). *ARIADNE (on line)*(40).

Motik, B., Patel-Schneider, P. F., & Parsia, B. (2009). OWL 2 Web Ontology Language. In C. Bock, A. Fokoue, P. Haase, R. Hoekstra, I. Horrocks, A. Ruttenberg, U. Sattler & M. Smith (Eds.): Consorcio World Wide Web (W3C). from http://www.w3.org/TR/owl2-syntax/. Retrieved.

Motz, R., Sosa, R., & Rodriguez, A. (2006). *Recycling Course Web Pages for the Semantic Web.* Paper presented at the LA-WEB '06: Proceedings of the Fourth Latin American Web Congress, 82-90.

Muramatsu, J., & Pratt, W. (2001). *Transparent Queries: investigation users' mental models of search engines.* Paper presented at the SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, 217-224.

Navigli, R., & Velardi, P. (2003). *An Analysis of Ontology-based Query Expansion Strategies.* Paper presented at the Workshop on Adaptive Text Extraction and Mining.

Neches, R., Fikes, R., Finin, T., Gruber, T., Patil, R., Senator, T., et al. (1991). Enabling technology for knowledge sharing. *AI Magazine, 12*(3), 36-56.

Nelson, S. J., Johnston, W. D., & Humphreys, B. L. (2001). Relationships in Medical Subject Headings (MeSH). In *Relationships in the Organization of Knowledge* (pp. 171-184): USA Kluwer Academic Publishers.

Nesbit, J., Belfer, K., & Vargo, J. (2002). A Convergent Participation Model for Evaluation of Learning Objects. *Canadian Journal of Learning and Technology, 28*(3), 105-120.

Neven, F., & Duval, E. (2002). *Reusable learning objects: a survey of LOM-based repositories.* Paper presented at the MULTIMEDIA '02: Proceedings of the tenth ACM international conference on Multimedia, 291-294.

Nielsen, M. L., & Ingwersen, P. (1999). *The Word Association Methodology - A Gateway to Work-Task Based Retrieval.* Paper presented at the Proceedings of Mira 99: Evaluating Interactive Information Retrieval.

Ochoa, X., & Duval, E. (2006). *Towards automatic evaluation of learning object metadata quality.* Paper presented at the In: Advances in Conceptual Modeling - Theory and Practice, ER 2006 Workshops BP-UML, CoMoGIS, COSS, ECDM, OIS, QoIS, SemWAT, 372-381.

Ochoa, X., & Duval, E. (2008). Relevance Ranking Metrics for Learning Objects. *IEEE Transactions on Learning Technologies, 1*(1), 34-48.

Ochoa, X., & Duval, E. (2009). Quantitative Analysis of Learning Object Repositories. *IEEE Transactions on Learning Technologies, 2*(3), 226-238.

OKI project. (2010). Open Knowledge Initiative. Retrieved Access ^, 2009, from http://www.okiproject.org/view/html/site/oki/node/382

Ouyang, Y., & Zhu, M. (2007). eLORM: Learning Object Relationship Mining based Repository. *E-Commerce Technology, IEEE International Conference on, and Enterprise Computing, E-Commerce, and E-Services, IEEE International Conference on, 0*, 691-698.

Paice, C. D. (1990). Another stemmer. *SIGIR Forum, 24*(3), 56-61.

Paulsson, F., & Naeve, A. (2006). *Establishing technical quality criteria for Learning Objects*. Paper presented at the eChallenges, 1431-1439.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, Automated library and information systems, 14*(3), 130-137.

Porter, M. F., & Sparck-Jones, K. W. P. (1997). *An algorithm for suffix stripping*: Morgan Kaufmann Publishers Inc.

Prieto M., M. E., Menéndez D., V. H., Segura N., A., & Vidal C., C. (2008). A Recommender System Architecture for Instructional Engineering In S. B. Heidelberg (Ed.), *Emerging Technologies and Information Systems for the Knowledge Society* (Vol. Volume 5288/2008, pp. 314-321). Berlin.

ProtegeWiki. (2010, 31 agosto 2010). Protege Ontology Library. Retrieved Access ^, 2010, from http://protegewiki.stanford.edu/wiki/Protege_Ontology_Library#OWL_ontologies

Qin, J., & Paling, S. (2001). Converting a controlled vocabulary into an ontology: the case of GEM. *Information Research, 6*(2).

Qin, J., Wang, H., & Shao, H. (2009). *Expansion Model of Semantic Query Based on Ontology*. Paper presented at the WMWA '09: Proceedings of the 2009 Second Pacific-Asia Conference on Web Mining and Web-based Application, 86-90.

Racer Systems GmbH, & Co. KG. (2009). Renamed Abox and Concept Expression Reasoner (Version RacerPro 2.0) [Ontology Reasoner]. on Line. Available http://www.racer-systems.com/

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics, v*(n), 17-30.

Reynolds, D. (2010, 28 Marzo). Jena 2 Inference support. Retrieved Access ^, 2010, from http://jena.sourceforge.net/inference/

Rich, E., & Knight, K. (1991). *Artificial Intelligence*: McGraw-Hill Science/Engineering/Math.

Rocchio, J. J., & Salton, G. (1971). *The SMART retrieval system - experiments in automatic document processing*: Prentice Hall.

Ruthven, I., & Lalmas, M. (2003). A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review, 18*(2), 95-145.

Salton, G., & McGill, M. J. (1986). *Introduction to Modern Information Retrieval*: McGraw-Hill, Inc.

Samper Zapater, J. J. (2006). *Ontologías para servicios web Semánticos de Información de Tráfico: Descripción y herramientas de explotación*. Universidad de Valencia.

Sangoi Pizzato, L. A., & Strube de Lima, V. L. (2003). *Evaluation of a thesaurus-based query expansion technique*. Paper presented at the PROPOR'03: Proceedings of the 6th international conference on Computational processing of the Portuguese language, 251-258.

Sarnowski, J., & Kessel, S. (2009). *Building standardized digital collections: ResCarta tools, a demo.* Paper presented at the Proceedings of the 13th European conference on Research and advanced technology for digital libraries, Corfu, Greece, 475-476

School of Computer Science- University of Manchester. (2010). FaCT OWL-DL reasoner, FaCT++ (Version 1.4.1) [Ontology Reasoner]. on Line. Available http://owl.man.ac.uk/factplusplus/

Segura N., A., Menéndez, V., & Prieto, M. (2008). Búsqueda y Composición de Objetos de Aprendizaje. In A. B. Gil González, J. A. Velázquez Iturbide & F. J. García Peñalvo (Eds.), *Proccedings of X International Symposium on Computers in Education SIIE* (pp. 335-340): Universidad de Salamanca, Ediciones Universidad de Salamanca.

Segura N., A., Vidal, C., Menéndez, V., Zapata, A., & Prieto, M. (2009). Exploring Characterizations of Learning Object Repositories Using Data Mining Techniques. In F. Sartori, M.-Á. Sicilia & N. Manouselis (Eds.), *Metadata and Semantic Research* (pp. 215-225): Springer-Verlag Berlin Heidelberg 2009.

Segura N., A., Vidal C., C., Campos G., P., Menéndez, V., & Prieto, M. (2011). Exploring characterizations of learning object repositories using data mining techniques - In press. *The Electronic Library, 29*(2).

Segura N., A., Vidal C., C., & Prieto, M. (2008). Evaluación de la Calidad del Software para el Aprendizaje. In A. B. Gil González, J. A. Velázquez Iturbide & F. J. García Peñalvo (Eds.), *Proccedings of X Simposio Internacional de Informática Educativa SIIE 2008* (pp. 59-64): Universidad de Salamanca, Ediciones Universidad de Salamanca.

Segura N., A., Vidal C., C., & Prieto M., M. (2010, 9 Julio 2010). *Query Expansion based on Domain Ontology for Learning Objects Search*. Paper presented at the 3rd IEEE International Conference on Computer Science and Information Technology, Chengdu, China.

Sicilia, M.-Á., García-Barriocanal, E., Pages, C., Martinez, J., & Gutierrez, J. (2005). Complete metadata records in learning object repositories some evidence and requirements. *International Journal of Learning Technology, 1*(4), 411-424.

Sicilia, M.-A., García-Barriocanal, E., Sánchez-Alonso, S., & Soto, J. (2005). *A Semantic Lifecycle Approach to Learning Object Repositories*. Paper presented at the AICT-SAPIR-ELETE '05: Proceedings of the Advanced Industrial Conference on Telecommunications/Service Assurance with Partial and Intermittent Resources Conference/E-Learning on Telecommunications Workshop, 466-471.

Sihvonen, A., & Vakkari, P. (2004a). Subject knowledge improves interactive query expansion assisted by a thesaurus. *Journal of Documentation, 60*(6).

Sihvonen, A., & Vakkari, P. (2004b). *Subject knowledge, thesaurus-assisted query expansion and search success*. Paper presented at the In Proceedings of the RIAO 2004 Conference, 393-404.

Simon, B., Massart, D., van Assche, F., Ternier, S., Duval, E., Brantner, S., et al. (2005). *A Simple Query Interface for Interoperable Learning Repositories*. Paper presented at the Proceedings of the 1st Workshop on Interoperability of Web-based Educational Systems, 11-18.

Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., & Katz, Y. (2007). Pellet: A practical OWL-DL reasoner. *Web Semant., 5*(2), 51-53.

Smith, L. A., Smith, E. T., & Melder, T. F. (2007). A taste of MERLOT: tutorial presentation. *Journal of Computing Sciences in Colleges, 22*(5), 147-148.

Song, M., Song, I.-Y., Hu, X., & Allen, R. (2007). Integration of association rules and ontologies for semantic query expansion. *Data & Knowledge Engineering, 63*(1), 63-75.

Soto, J., García-Barriocanal, E., & Sánchez-Alonso, S. (2007). Semantic learning object repositories. *International Journal of Continuing Engineering Education and Life-Long Learning (IJCEELL), 17*, 432-446.

Spearman, C. (1987). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology, 100*(3/4), 441-471.

Spink, A., Greisdorf, H., & Bateman, J. (1998). From highly relevant to not relevant: examining different regions of relevance. *Information Processing and Management, 34*(5), 599-621.

Spink, A., Jansen, B. J., Wolfram, D., & Saracevic, T. (2002). From E-Sex to E-Commerce: Web Search Changes. *Computer, 35*(3), 107-109.

Spink, A., Wolfram, D., Jansen, M., & Saracevic, T. (2001). Searching the Web: the public and their queries. *Journal of the American Society for Information Science and Technology, 52*(3), 226-234.

Stanford Center for Biomedical Informatics Research at the Stanford University School of Medicine. (2010). The Protégé Ontology Editor and Knowledge Acquisition System (Version 4.0.2) [Ontology Editor]. on Line. Available http://protege.stanford.edu/

Stefaner, M., Vecchia, E. D., Condotta, M., Wolpers, M., Specht, M., Apelt, S., et al. (2007). *MACE - Enriching Architectural Learning Objects for Experience Multiplication.* Paper presented at the EC-TEL, 322-336.

Steve, G., Gangemi, A., Pisanelli, D. M., Charrell, P. J., Kangassalo, H., & Jaakkola, H. (1998). Integrating Medical Terminologies with ONIONS Methodology. In *Information Modelling and Knowledge Bases IX* (Vol. 45, pp. 1-18): IOS-Press.

Stojanovic, N. (2005). *An Approach for Defining Relevance in the Ontology-Based Information Retrieval.* Paper presented at the WI '05: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, 359-365.

Stojanovic, N., Studer, R., & Stojanovic, L. (2004). *An Approach for Step-By-Step Query Refinement in the Ontology-Based Information Retrieval.* Paper presented at the WI '04: Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence, 36-43.

Tello Cáceres, J. (2007). *Estudio exploratorio de defectos en registros de meta-datos IEEE LOM de objetos de aprendizaje.* Paper presented at the Post-Proceedings del IV Simposio Pluridisciplinar sobre DiseÃ±o, EvaluaciÃ³n y Desarrollo de Contenidos Educativos Reutilizables, SPDECE 2007.

Ternier, S., Bosman, B., Duval, E., Metzger, L., Halm, M., Thorne, S., et al. (2006). *Connecting OKI And SQI: One Small Piece Of Code, A Giant Leap For Reusing Learning Objects.* Paper presented at the Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2006, 825-831.

Ternier, S., Massart, D., Campi, A., Guinea, S., Ceri, S., & Duval, E. (2008). Interoperability for searching learning object repositories: The proLearn query language. *D-Lib Magazine, 14*(1-2).

Ternier, S., Olmedilla, D., & Duval, E. (2005). *Peer-to-Peer versus Federated Search: towards more Interoperable Learning Object Repositories.* Paper presented at the Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2005, 1421-1428.

Tuominen, J., Kauppinen, T., Viljanen, K., & Hyonen, E. (2009). *Ontology-Based Query Expansion Widget for Information Retrieval.* Paper presented at the

Proceedings of the 5th Workshop on Scripting and Development for the Semantic Web (SFSW 2009), 6th European Semantic Web Conference (ESWC 2009).

Uebersax, J. S. (1983). A design-independent method for measuring the reliability of psychiatric diagnosis. *Journal on Psychiatric Research, 17*(4), 335-342.

Uschold, M., Benjamins, V. R., Gómez-Pérez, B. C. A., Guarino, N., & Robert, J. (1999). *A framework for understanding and classifying ontology applications.* Paper presented at the in Proceedings of the IJCAI99 Workshop on Ontologies and Workshop on Ontologies and Problem-Solving Methods: Lessons Learned and Future Trends. (IJCAI99).

Vakkari, P. (2002). *Subject Knowledge, Source of Terms, and Term Selection in Query Expansion: An Analytical Study.* Paper presented at the Proceedings of the 24th BCS-IRSG European Colloquium on IR Research, 110-123.

Van de Sompel, H., & Lagoze, C. (2002). *Notes from the Interoperability Front:A Progress Report on the Open Archives Initiative.* Paper presented at the European Conference on Digital Libraries.

Van Heijst, G., Schreiber, A. T., & Wielinga, B. J. (1997). Using explicit ontologies in KBS development. *International Journal of Human-Computer Studies, 46*(2-3).

Vargo, J., Nesbit, J. C., Belfer, K., & Archambault, A. (2003). Learning Object Evaluation: Computer-Mediated Collaboration and Inter-Rater Reliability. *International Journal of Computers and Application, 25*, 198-205.

Vidal, C., Segura N., A., & Prieto, M. (2008). *Calidad en objetos de aprendizaje.* Paper presented at the V Simposio Pluridisciplinar sobre Diseño y Evaluación de Contenidos Educativos Reutilizables.

Vidal C., C., Segura N., A., Campos G., P., & Sánchez, S. (2010). *Quality in Learning Objects: Evaluating compliance with metadata Standards.* Paper presented at the Fourth International Conference on Metadata and Semantics Research.

Voorhees, E. M. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management: an International, 36*(5), 697-716.

Voorhees, E. M. (2001). *Evaluation by highly relevant documents.* Paper presented at the SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, 74-82.

W3C. (2008). Guía Breve de Web Semántica. on Line: Consorcio World Wide Web (W3C). from http://www.w3c.es/divulgacion/guiasbreves/websemantica. Retrieved Septiembre 2010.

Wang, P.-H., Wang, J.-Y., & Lee, H.-M. (2004). *QueryFind: Search Ranking Based on Users' Feedback and Expert's Agreement.* Paper presented at the EEE '04:

Proceedings of the 2004 IEEE International Conference on e-Technology, e-Commerce and e-Service (EEE'04), 299-304.

WASP Center Server. (2006). Multi-version Ontology REasoner (Version 0.8.0 for Windows2000/WindowsXP) [Ontology Reasoner]. on Line. Available http://wasp.cs.vu.nl/sekt/more/

Weigand, H. (1997). *Multilingual ontology-based lexicon for News Filtering.* Paper presented at the Proc. IJCAI workshop on multilingual ontologies.

Wenying, G., & Deren, C. (2006). *Semantic Approach for e-learning System.* Paper presented at the IMSCCS '06: Proceedings of the First International Multi-Symposiums on Computer and Computational Sciences - Volume 2 (IMSCCS'06), 442-446.

Wiley, D. A. (2003). Learning objects: difficulties and opportunities [Electronic Version]. Retrieved Marzo 2010 from http://reusability.org/read/chapters/wiley.doc.

Winston, P. H. (1992). *Artificial Intelligence (Tercera Edición)*: Addison-Wesley.

Wollersheim, D., & Rahayu, W. J. (2005). *Using Medical Test Collection Relevance Judgements to Identify Ontological Relationships Useful for Query Expansion.* Paper presented at the ICDEW '05: Proceedings of the 21st International Conference on Data Engineering Workshops.

Xiang, X., Shen, Z., Guo, L., & Shi, Y. (2003). Introduction of the Core Elements Set in Localized LOM Model. *IEEE Learning Technology Task Force, 5*(1).

Yue, W., Chen, Z., Lu, X., Lin, F., & Liu, J. (2005). *Using Query Expansion and Classification for Information Retrieval.* Paper presented at the SKG '05: Proceedings of the First International Conference on Semantics, Knowledge and Grid.

Zazo, Á. F., Figuerola, C. G., Alonso Berrocal, J. L., & Emilio, R. (2005). Reformulation of queries using similarity thesauri. *Information Processing and Management: an International Journal, 41*(5), 1163-1173.

Zhang, B., Du, Y., Li, H., & Jia, L. (2009). The Method of Query Expansion Based on Domain Ontology. *Circuits, Communications and Systems, Pacific-Asia Conference on, 0*, 755-758.

Zhong, M., Chen, Z., Lin, Y., & Yao, J. (2004). *Using classification and key phrase extraction for information retrieval.* Paper presented at the Proceedings of the World Congress on Intelligent Control and Automation (WCICA), 3037-3041.

Zou, G., Zhang, B., Gan, Y., & Zhang, J. (2008). *An Ontology-Based Methodology for Semantic Expansion Search.* Paper presented at the FSKD '08: Proceedings of the 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 453-457.

Anexos

Anexo A: Resumen de resultados obtenidos desde el repositorio para el experimento

Consulta		original	Concept ocercano	Tipos de expansión								
				syn_exa	*syn_all*	*isa_exa*	*isa_her*	*isa_hij*	*par_exa*	*par_otr*	*reg_a*	*reg_por*
T1	population genetic	1	-	-	-	-	-	-	-	-	-	-
T2	mutation	10	0	0	-	0	0	0	-	-	-	-
T3	DNA replication	10	-	-	0	0	42	20	-	-	-	0
T4	gene expression	8	-	-	-	0	0	-	-	-	-	1
T5	recombination	10	1	-	-	0	30	10	-	-	-	0
T6	chromosome	10	-	-	0	2	18	0	-	-	-	-
T7	DNA	10	-	-	-	-	-	-	-	-	-	-
T8	mendelian genetic	2	-	-	-	-	-	-	-	-	-	-
T9	transposable element	0	-	-	-	-	-	-	-	-	-	-
T10	linkage	10	10	-	-	10	30	40	-	-	-	-
T11	molecular genetic	2	0	0	-	0	0	0	-	-	-	-
T12	biotechnology	10	-	-	-	-	-	-	-	-	-	-
T13	developmental genetic	0	-	-	-	-	-	-	-	-	-	-
T14	genetic	10	0	-	0	10	19	10	-	-	-	-
T15	meiosis	10	-	-	-	0	30	0	0	-	-	0
T16	transcription	10	-	-	-	0	27	0	8	13	-	0
T17	translation	10	-	12	-	0	20	0	8	13	-	1
T18	ADH in drosophila	0	-	-	-	-	-	-	-	-	-	-
T19	cytogenetic	1	-	-	-	-	-	-	-	-	-	-
T20	genetic of cancer	0	-	-	-	-	-	-	-	-	-	-
T21	genomic	10	-	-	-	-	-	-	-	-	-	-
T22	introduction	10	-	-	-	-	-	-	-	-	-	-
T23	mitosis	10	-	-	-	0	30	0	0	0	-	0
T24	conjugation	10	-	-	-	10	19	0	-	-	-	0
T25	DNA structure	10	0	-	-	10	40	20	-	-	-	-
T26	gene mapping	0	-	-	-	-	-	-	-	-	-	-
T27	gene regulation	4	0	-	-	0	0	0	0	0	0	-
T28	genetic variation	4	-	-	-	-	-	-	-	-	-	-
T29	genetic of bacteria	0	-	-	-	-	-	-	-	-	-	-
T30	inheritance	10	0	0	-	0	0	-	-	-	-	-
T31	migration	10	0	-	-	1	0	4	-	-	-	0
T32	non-synonymous	0	-	-	-	-	-	-	-	-	-	-

Consulta		original	Concept ocercano	Tipos de expansión								
				syn_exa	*syn_all*	*isa_exa*	*isa_her*	*isa_hij*	*par_exa*	*par_otr*	*reg_a*	*reg_por*
	substitution											
T33	quantitative genetic	0	-	-	-	-	-	-	-	-	-	-
T34	quantitative trait	1	-	-	-	-	-	-	-	-	-	-
T35	transmission genetic	0	-	-	-	-	-	-	-	-	-	-
T36	bacterial genetic	0	-	-	-	-	-	-	-	-	-	-
T37	bacteriophage	1	-	-	-	-	-	-	-	-	-	-
T38	bioinformatic	10	-	-	-	-	-	-	-	-	-	-
T39	CatLab	0	-	-	-	-	-	-	-	-	-	-
T40	cell cycle	10	-	10	-	0	11	0	-	-	-	0
T41	cell cycle control	0	0	-	0	0	0	-	0	0	-	-
T42	cell cycle regulation	0	0	0	0	0	10	0	10	0	10	-
T43	centromeres	1	-	-	-	-	-	-	-	-	-	-
T44	cloning	10	-	-	-	-	-	-	-	-	-	-
T45	disease	10	0	-	-	-	-	-	-	-	-	-
T46	DNA repair	3	-	-	-	0	27	29	-	-	-	0
T47	DNA technology	2	-	-	-	-	-	-	-	-	-	-
T48	DNA transcription	1	0	-	-	10	0	0	-	-	-	0
T49	evolution	10	-	-	-	-	-	-	-	-	-	-
T50	evolutionary inference	0	-	-	-	-	-	-	-	-	-	-
T51	gene	10	0	0	0	0	0	-	-	-	-	-
T52	gene interaction	2	0	0	-	0	0	0	-	-	-	-
T53	gene isolation	0	-	-	-	-	-	-	-	-	-	-
T54	gene manipulation	0	-	-	-	-	-	-	-	-	-	-
T55	gene mapping in eukaryotes	0	-	-	-	-	-	-	-	-	-	-
T56	gene therapy	9	-	-	-	-	-	-	-	-	-	-
T57	genetic of virus	0	-	-	-	-	-	-	-	-	-	-
T58	genotype structure of population	0	-	-	-	-	-	-	-	-	-	-
T59	granite outcrop	0	-	-	-	-	-	-	-	-	-	-
T60	hardy weinberg principle	0	-	-	-	-	-	-	-	-	-	-
T61	hemoglobinopathy	0	-	-	-	-	-	-	-	-	-	-
T62	heredity	10	-	-	-	-	-	-	-	-	-	-
T63	human disease	5	-	-	-	-	-	-	-	-	-	-
T64	incompatibility	1	0	-	-	-	-	-	-	-	-	-
T65	introduction to DNA	3	-	-	-	-	-	-	-	-	-	-

Consulta		original	Concept ocercano	Tipos de expansión								
				syn_exa	*syn_all*	*isa_exa*	*isa_her*	*isa_hij*	*par_exa*	*par_otr*	*reg_a*	*reg_por*
T66	linkage of gene	0	-	-	-	-	-	-	-	-	-	-
T67	mendel law	0	-	-	-	-	-	-	-	-	-	-
T68	mendelian inheritance	4	-	-	-	-	-	-	-	-	-	-
T69	molecular evolution	0	-	-	-	-	-	-	-	-	-	-
T70	mutation effect	0	-	-	-	-	-	-	-	-	-	-
T71	mutation repair	0	-	-	-	-	-	-	-	-	-	-
T72	non-mendelian inheritance	0	-	-	-	-	-	-	-	-	-	-
T73	PCR	10	-	-	-	-	-	-	-	-	-	-
T74	pedigree	1	-	-	-	-	-	-	-	-	-	-
T75	phenotype structure of population	0	-	-	-	-	-	-	-	-	-	-
T76	plasmid	3	0	-	-	-	-	-	-	-	-	-
T77	protein	10	-	-	-	-	-	-	-	-	-	-
T78	random drift	0	-	-	-	-	-	-	-	-	-	-
T79	recombinant DNA technique	1	-	-	-	-	-	-	-	-	-	-
T80	recombinant DNA	10	-	-	-	-	-	-	-	-	-	-
T81	recombinant DNA technology	2	-	-	-	-	-	-	-	-	-	-
T82	RNA	10	-	-	-	-	-	-	-	-	-	-
T83	RNA processing	2	-	-	-	0	0	3	8	21	-	-
T84	sex	10	0	-	-	0	1	-	0	0	-	-
T85	sex determination	2	-	-	-	0	0	2	-	-	-	-
T86	sex linkage	1	-	-	-	-	-	-	-	-	-	-
T87	stem cell	10	0	0	-	10	8	2	-	-	-	-
T88	variation in chromosome number	0	-	-	-	-	-	-	-	-	-	-
T89	variation in chromosome structure	0	-	-	-	-	-	-	-	-	-	-
T90	viral genetic	0	-	-	-	-	-	-	-	-	-	-
T91	WRRB	0	-	-	-	-	-	-	-	-	-	-

Anexo B : Lista de resultados ponderados, caso de estudio

Q	tipo	OA	título	p1	p2	p3	p4	veces	mediana	po	rank
T10	*isa_her*	91119	Protein Explorer	0	0	0		3	0	0,00	1
		90984	Biochemistry in 3D-Lehninger Principles of Biochemistry	1	1	1		3	1	0,33	2
		85278	Transcription (DNA to RNA)	2	2	2		3	2	0,67	3
		85279	The Genetic Code	3	3	3		3	3	1,00	4
		85280	Significance of DNA Sequence	4	4	4		3	4	1,33	5
		85281	RNA Processing (post-transcriptional modifications)	5	5	5		3	5	1,67	6
		85260	One Gene, One Protein	6	6	6		3	6	2,00	7
		79210	Molecular Biochemistry	7	7	7		3	7	2,33	8
		90940	Blast	8	8	8		3	8	2,67	9
		243074	The Genetic Code	9	9	9		3	9	3,00	10
	isa_hij	91119	Protein Explorer	0	0	0	0	4	0	0,00	1
		90984	Biochemistry in 3D-Lehninger Principles of Biochemistry	1	1	1	1	4	1	0,25	2
		85278	Transcription (DNA to RNA)	2	2	2	2	4	2	0,50	3
		85279	The Genetic Code	3	3	3	3	4	3	0,75	4
		85280	Significance of DNA Sequence	4	4	4	4	4	4	1,00	5
		85281	RNA Processing (post-transcriptional modifications)	5	5	5	5	4	5	1,25	6
		85260	One Gene, One Protein	6	6	6	6	4	6	1,50	7
		79210	Molecular Biochemistry	7	7	7	7	4	7	1,75	8
		90940	Blast	8	8	8	8	4	8	2,00	9
		243074	The Genetic Code	9	9	9	9	4	9	2,25	10
T15	*isa_her*	91401	The Biology Project: Cell Biology	0	0			2	0	0,00	2
		90943	Biological Basis of Heredity	3	1			2	2	1,00	5
		77749	Mitosis & Meiosis	8	5	7		3	7	2,33	9
		91398	Cells Alive!	6	4			2	5	2,50	10
T16	*par_otr*	85281	RNA Processing (post-transcriptional modifications)	2	0			2	1	0,50	3
T17	*isa_her*	79615	Biology Flash Animations	4	4			2	4	2,00	6

Q	tipo	OA	título	p1	p2	p3	p4	veces	mediana	po	rank
	par_otr	85281	RNA Processing (post-transcriptional modifications)	1	0			2	0,5	0,25	3
T23	*isa_her*	91401	The Biology Project: Cell Biology	0	0			2	0	0,00	2
		90943	Biological Basis of Heredity	1	2			2	1,5	0,75	3
		75251	Meiosis Animation	2	3			2	2,5	1,25	6
		81501	Home made genetics animations	3	4			2	3,5	1,75	7
		77749	Mitosis & Meiosis	5	7	6		3	6	2,00	8
		91398	Cells Alive!	4	5			2	4,5	2,25	10
		86024	Lew-Port's Biology Place--Animated Reviews	6	7			2	6,5	3,25	12
		80042	Biology Animation	7	8			2	7,5	3,75	13
		82020	The Biology Place	8	9			2	8,5	4,25	15
T25	*isa_her*	80336	Cash Flow Statement	0	1			2	0,5	0,25	4
		84788	Business Transactions Tutorial	7	2			2	4,5	2,25	11
		84814	Periodic Inventory	8	3			2	5,5	2,75	12
	isa_hij	199783	Drills for Accounting Basics	0	4			2	2	1,00	2
T3	*isa_her*	85217	Kimball's Biology Pages	1	0			2	0,5	0,25	7
		85281	RNA Processing (post-transcriptional modifications)	1	2			2	1,5	0,75	8
	isa_hij	85278	Transcription (DNA to RNA)	3	0			2	1,5	0,75	2
		85282	RNA Viruses	8	1			2	4,5	2,25	6
T46	*isa_hij*	358537	Mouse Party	0	0			2	0	0,00	1
		89967	DNA Structure 1.0	0	1			2	0,5	0,25	4
T83	*par_otr*	85281	RNA Processing (post-transcriptional modifications)	1	2			2	1,5	0,75	4

Anexo C : Instrumento para la evaluación de la relevancia de los resultados

¿El recurso para el Aprendizaje es relavante o satisface la consulta aplicada?. La evaluación debe ser realizada en 3 niveles: (N) no relevante, (PR) Parcialmente Relevante, (R) Relevante

RECURSO PARA EL APRENDIZAJE			EVALUACION	
TÍTULO	DESCRIPCION	Ver Recurso	CONSULTA	¿el recurso satisface la consulta?
Flash animations in science	Learning object repository for science teachers (biology, physics, mathematics).You will find flash animations and simulations especially made by teachers for teachers.The main topics are: Life sciences (transcription, replication, DNA, genetics, anatomy, organs, heart cycle, digestive tract, physiology) Earth and space sciences (astronomy, geophysics, satellisation, seismic wave,plates, continental drift, tsunami, resources and environment, water cycle) Electromagnetism (electric field, potential, magnetic field, magnet and compass, charge, Faraday's law of induction) Optics (light, shadow, lens, geometrical optic, physical optic, laser, waves) Mechanics (motion and force, Newton's law, gravitation, energy, nuclear)	http://www.edumedia-sciences.com	transcription	
eStroke Animated Chinese Characters	The eStroke, site provides an, ideal tool for learning Chinese language and character writing. Every, Chinese, character, is animated, stroke by stroke in an appropriate speed. The characters are presented in both traditional and simplified forms as well as their corresponding, pinyin (phonetic spelling) and English translation. Users can also create new Chinese characters with, eContour in, a related, site.	http://www.eon.com.hk/estroke/	transcription	
			translation	
mRNA Processing--Tutorial and Animation	This site is part of the Virtual Cell Project at North Dakota State University. The site consists of three parts: an overview of mRNA processing and its key steps, an in-depth look at the major players and events involved in processing mRNA, and an embedded Windows Media movie version of the mRNA processing animation.	http://vcell.ndsu.nodak.edu/animations/mrnaprocessing/index.htm	transcription	
			translation	
DNA Replication Video	Video clip of cell wrapping and DNA replication on YouTube.	http://www.youtube.com/watch?v=Pj9cdVeIntY	translation	
			DNA replication	

Nota. En este anexo solo se han incluido 4 OA para ejemplificar el intrumento de evaluación utilizado por los expertos.

Anexo D : Extracto de la Implementación del prototipo

Class Expansion_Ontologia

java.lang.Object
 Expansion_Ontologia

---

public class Expansion_Ontologia
extends java.lang.Object

Field Summary	
static java.io.FileOutputStream	fout
static com.hp.hpl.jena.rdf.model.Model	model
static com.hp.hpl.jena.ontology.OntModel	ontModel
static java.io.OutputStreamWriter	out
static java.lang.String	prefijo_ontologia
static java.lang.String	prefijo_sparql

Constructor Summary
Expansion_Ontologia()

Method Summary	
static java.lang.String	busca_clase_con_label(java.lang.String⟨⟩term_buscado)
static java.lang.String	busca_ex_cod(java.lang.String clase)
static java.lang.String	busca_ex_label(java.lang.String buscado)
static void	carga(java.lang.String owlFile)
static boolean	existe_ap_label(java.lang.String buscado)

static boolean	existe_ex_label(java.lang.String buscado)
static java.util.ArrayList<java.lang .String>	label_synonyms_all(java.lang.String clase)
static java.util.ArrayList<java.lang .String>	label_synonyms_exacto(java.lang.String clase)
static void	lista_label()
static void	main(java.lang.String〈〉 args)
static java.util.ArrayList<java.lang .String>	sparql_al_que_regulate(java.lang.String clase)
static java.util.ArrayList<java.lang .String>	sparql_clase_que_la_reemplaza(java.lang.String clase)
static java.util.ArrayList<java.lang .String>	sparql_clase_quela_regulate(java.lang.String clase)
static java.util.ArrayList<java.lang .String>	sparql_hermano_is_a(java.lang.String clase)
static java.util.ArrayList<java.lang .String>	sparql_hermano_partof(java.lang.String clase)
static java.util.ArrayList<java.lang .String>	sparql_hijos_is_a(java.lang.String clase)
static void	sparql_ini()
static java.util.ArrayList<java.lang .String>	sparql_is_a_label(java.lang.String clase)
static java.util.ArrayList<java.lang .String>	sparql_is_a(java.lang.String clase)
static void	sparql_literales()
static void	sparql_OBONamespace(java.lang.String clase)
static java.util.ArrayList<java.lang .String>	sparql_partof_label(java.lang.String clase)
static java.util.ArrayList<java.lang .String>	sparql_partof(java.lang.String clase)

static int	sparql_quita_obsoleto(java.lang.String clase)
static void	sparql_regulates_all()

Methods inherited from class java.lang.Object
clone, equals, finalize, getClass, hashCode, notify, notifyAll, toString, wait, wait, wait

Field Detail

prefijo_ontologia
public static java.lang.String prefijo_ontologia

---

prefijo_sparql
public static java.lang.String prefijo_sparql

---

model
public static com.hp.hpl.jena.rdf.model.Model model

---

ontModel
public static com.hp.hpl.jena.ontology.OntModel ontModel

---

fout
public static java.io.FileOutputStream fout

---

out
public static java.io.OutputStreamWriter out

Constructor Detail
Expansion_Ontologia

public Expansion_Ontologia()
Method Detail
sparql_ini
public static void sparql_ini()

---

carga
public static void carga(java.lang.String owlFile)

---

busca_clase_con_label
public static java.lang.String busca_clase_con_label(java.lang.String〈〉 term_buscado)

---

lista_label
public static void lista_label()

---

label_synonyms_exacto

public static java.util.ArrayList<java.lang.String> label_synonyms_exacto(java.lang.String clase)

---

label_synonyms_all
public static java.util.ArrayList<java.lang.String> label_synonyms_all(java.lang.String clase)

---

sparql_partof
public static java.util.ArrayList<java.lang.String> sparql_partof(java.lang.String clase)

---

sparql_hermano_partof
public static java.util.ArrayList<java.lang.String> sparql_hermano_partof(java.lang.String clase)

---

sparql_partof_label
public static java.util.ArrayList<java.lang.String> sparql_partof_label(java.lang.String clase)

---

sparql_hermano_is_a
public static java.util.ArrayList<java.lang.String> sparql_hermano_is_a(java.lang.String clase)

---

sparql_is_a
public static java.util.ArrayList<java.lang.String> sparql_is_a(java.lang.String clase)

---

sparql_is_a_label
public static java.util.ArrayList<java.lang.String> sparql_is_a_label(java.lang.String clase)

---

sparql_hijos_is_a
public static java.util.ArrayList<java.lang.String> sparql_hijos_is_a(java.lang.String clase)

---

existe_ap_label
public static boolean existe_ap_label(java.lang.String buscado)

---

existe_ex_label
public static boolean existe_ex_label(java.lang.String buscado)

---

busca_ex_label
public static java.lang.String busca_ex_label(java.lang.String buscado)

---

busca_ex_cod
public static java.lang.String busca_ex_cod(java.lang.String clase)

---

sparql_OBONamespace
public static void sparql_OBONamespace(java.lang.String clase)

---

sparql_al_que_regulate
public static java.util.ArrayList<java.lang.String> sparql_al_que_regulate(java.lang.String clase)

---

sparql_clase_quela_regulate

public static java.util.ArrayList<java.lang.String> sparql_clase_quela_regulate(java.lang.String clase)

---

sparql_clase_que_la_reemplaza
public static java.util.ArrayList<java.lang.String> sparql_clase_que_la_reemplaza(java.lang.String clase)

---

sparql_quita_obsoleto
public static int sparql_quita_obsoleto(java.lang.String clase)

---

sparql_regulates_all
public static void sparql_regulates_all()

---

sparql_literales
public static void sparql_literales()

---

main
public static void main(java.lang.String〈〉 args)