UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA

ESCUELA DE CIENCIAS DE LA COMPUTACIÓN

COMPARACIÓN DE HERRAMIENTAS DE BÚSQUEDA PARA RECURSOS EDUCATIVOS ABIERTOS

TESIS PREVIA A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN SISTEMAS INFORMÁTICOS Y COMPUTACIÓN

Autor:

Israel Alejandro Cueva Hidalgo

Director:

Ing. Janneth Chicaiza Espinosa

Loja – Ecuador

2011

CESIÓN DE DERECHOS

Yo, ***Israel Alejandro Cueva Hidalgo****, declaro ser autor del presente trabajo y eximo expresamente a la Universidad Técnica Particular de Loja y a sus representantes legales de posibles reclamos o acciones legales.*

Adicionalmente declaro conocer y aceptar la disposición del Art. 67 del Estatuto Orgánico de la Universidad Técnica Particular de Loja que su parte pertinente textualmente dice: "Forman parte del parte del patrimonio de la Universidad la propiedad intelectual de investigaciones, trabajos científicos o técnicos y tesis de grado que se realicen a través, o con el apoyo financiero, académico o institucional (operativo) de la universidad".

.....

Israel A. Cueva H.

CERTIFICACIÓN

Ing.

Janneth Chicaiza Espinosa

DIRECTOR DE TESIS

CERTIFICA:

Haber dirigido y supervisado el desarrollo del presente proyecto de tesis previo a la obtención del título de ***INGENIERO EN SISTEMAS INFORMÁTICOS Y COMPUTACIÓN****, y una vez que este cumple con todas las exigencias y los requisitos legales establecidos por la Universidad Técnica Particular de Loja, autoriza su presentación para los fines legales pertinentes.*

Loja, 23 Marzo 2011

Ing. Janneth Chicaiza Espinosa

AUTORÍA

El presente proyecto de tesis con cada una de sus observaciones, análisis, evaluaciones, conclusiones y recomendaciones emitidas, es de absoluta responsabilidad del autor.

Además, es necesario indicar que la información de otros autores empleada en el presente trabajo está debidamente especificada en fuentes de referencia y apartados bibliográficos.

.....

Israel A. Cueva Hidalgo.

DEDICATORIA

A mis padres y abuelitos, por su apoyo y ayuda incondicional.

Israel Alejandro

AGRADECIMIENTOS

Mis sinceros agradecimientos a las personas que de una u otra manera me han apoyado para poder culminar esta tesis, a mis padres y mis abuelos, por su paciencia y esfuerzo en todo este tiempo, a mis compañeros de curso, por estos últimos años de apoyo y amistad incondicional, y a mi Directora de Tesis, Janneth Chicaiza, quién fue la guía durante este duro camino que finalmente da sus frutos.

Israel Alejandro

Índice General

1. Estudio de Sistemas de Búsqueda para Recursos Educativos Abiertos 1

Propósito 2

Introducción 2

1.1 Introducción a Open Educational Resources (OER) 4

1.1.1 Definición de OER 4

1.1.2 Ventajas y Desventajas de los OER 4

1.1.3 Licenciamiento 5

1.1.4 Repositorios para Recursos Educativos Abiertos 6

1.1.5 Estado Actual – Iniciativa de OER 6

1.2 Sistemas de Búsqueda 7

1.2.1 Definiciones 9

1.2.2 Proceso de Recuperación de Información 9

1.2.3 Elementos Clave de las Búsquedas 10

1.2.4 Técnicas de Recuperación de Información 11

1.2.5 Modelos de Recuperación de Información 12

1.2.6 Arquitectura General de un Sistema de Recuperación 12

1.2.7 Búsqueda Sobre Datos Estructurados 14

1.3 Búsqueda de OER 17

1.3.1 Uso de Buscadores Tradicionales 17

1.3.2 Uso de Buscadores Especializados 18

1.3.3 Uso de Buscadores por Tipo de Recurso 18

1.3.4 Estándares de Metadatos de Material Educativo 19

1.4 Herramientas de Búsqueda 20

1.4.1 Courseware Watchdog 20

1.4.2 Knowledge Hub 20

1.4.3 DiscoverEd 21

1.4.4 Lucene 21

1.4.5 SOLR 21

1.4.6 SIREn 22

1.4.7 Regain 22

1.4.8 Nutch 22

1.5 Comparación y Selección de las Herramientas de Búsqueda 23

2. Evaluación del Sistema de Búsqueda 26

Propósito 27
Introducción 27
2.1 Definiciones 28
2.1.1 Relación entre Precisión y Exhaustividad 29
2.1.2 Cálculo de la Precisión y de la Exhaustividad 29
2.2 Evaluación a Nivel de Usuario 31
2.2.1 Interfaz de Usuario 31
2.2.2 Sistema de Interrogación e Interfaz de Búsqueda 32
2.2.3 Recuperación y Consulta 32
2.2.4 Evaluación de los Sistemas de Recuperación de Información 33
2.2.5 Características Generales del Sistema de Búsqueda para OERs 34
2.2.6 Resumen de Criterios y Aspectos para la Evaluación a Nivel de Usuario 34
2.3 Evaluación del Rendimiento 35
3. Análisis y Selección de Fuentes de Material Educativo para la Búsqueda 40
Propósito 41
Introducción 41
3.1 Búsqueda y Evaluación de los Recursos Web 42
3.1.1 Indicadores Clave de la Calidad 42
3.1.2 Criterios de Calidad y Evaluación de la información en Internet 43
3.2 Criterios a Considerarse para la Selección de los Repositorios 47
4. Implementación de las Herramientas de Búsqueda 51
Propósito 52
Introducción 52
4.1 Esquema de Recuperación para los Repositorios de Material Educativo Seleccionados 53
4.2 Implementación de Herramientas 53
4.2.1 Características Hardware y Software 53
4.2.2 Instalación y Configuración de las Herramientas 54
4.3 Resultados Preliminares 57
4.4 Discusión 59
5. Comparación de las Herramientas de Búsqueda para Material Educativo 61
Propósito 62
Introducción 62
5.1 Evaluación de Criterios a Nivel de Usuario 63
5.2 Evaluación del Rendimiento 67

6. Conclusiones y Recomendaciones 77
6.1 Conclusiones 78
6.2 Recomendaciones 79
7. Bibliografía 80
8. Anexos 84
8.1 Anexo 1 - Estudio de las Herramientas Seleccionadas 86
8.1.1 Lucene 86
8.1.2 Regain 88
8.1.3 Nutch 92
8.2 Anexo 2 - Instalación de Nutch 96
8.2.1 Software Necesario 96
8.2.2 Configuración de Nutch 1.2 97
8.3 Anexo 3 - Instalación Regain 103
8.4 Anexo 4 - Disponibilidad de Cursos en los Repositorios Seleccionados 106
8.5 Anexo 5 - Archivos Disponibles para la Categoría “Anthropology” en cada Repositorio. 110
8.6 Anexo 6 – Configuración del Archivo ‘otech_crawl.txt’ -Enlaces Web de la Categoría Antropología de los Repositorios 126
8.7 Anexo 7 - Configuración del Archivo ‘crawl-urlfilter.txt’ de Nutch. 127
8.8 Anexo 8 – Ejemplo: Relación entre Exhaustividad y Precisión. 131
8.9 Anexo 9 – Encuesta al Usuario 132
8.9.1 Requerimientos de los SRI Propuestos 132

Índice de Figuras

Figura 1.1Tipos de OERs (Hewlett, 2005). 4

Figura 1.2 Proceso de Recuperación de Información (Salazar & Pinto, 2003). 10

Figura 1.3 Técnicas de recuperación de información (Pinto, 2009). 11

Figura 1.4 Clasificación de los SRI (López, 2006) (Martínez F. J., 2002). 12

Figura 1.5 Arquitectura general de un SRI (Mel'nikov et al., 2008). 13

Figura 2.1 Evolución típica de la precisión y de la exhaustividad en un SRI (Martínez F. J., 2002). 29

Figura 2.2 Cálculo de pares de valores E-P de la búsqueda de ejemplo (Martínez F. J., 2002). 30

Figura 2.3 Proceso para evaluar el rendimiento del sistema de búsqueda. 39

Figura 4.1 Esquema de recuperación para recuperar OERs desde distintos repositorios. 53

Figura 4.2 Utilización de plugins disponibles en Nutch 1.2 55

Figura 4.3 Configuración mimeType: "rdf+xml" 55

Figura 4.4 Configuración mimeType: "xhtml+xml" 56

Figura 4.5 Configuración mimeType: "rss+xml" 56

Figura 4.6 Configuración mimeType: "text+xml" 56

Figura 4.7 Configuración del archivo 'crawl-urlfilter'. 57

Figura 4.8 Configuración del archivo 'crawl-urlfilter' - delimitador 57

Figura 4.9 Búsqueda utilizando la herramienta Nutch. 58

Figura 4.10 Búsqueda utilizando la herramienta Nutch (Muestra 2). 58

Figura 4.11 Búsqueda utilizando la herramienta Regain. 59

Figura 4.12 Búsqueda utilizando la herramienta Regain (Muestra 2). 59

Figura 5.1 Comparación de calificaciones para Nutch 66

Figura 5.2 Comparación de calificaciones para Regain 67

Figura 5.3 Realización de una consulta en el buscador Nutch. 70

Figura 5.4 Realización de una consulta en el buscador Regain. 71

Figura 8.1 Comparación de las variantes de Regain (Schneider & Tesche, Regain manual, 2009). 88

Figura 8.2 Arquitectura de Nutch (Cutting, 2004). 93

Figura 8.3 Comparación entre las versiones de Tomcat. 96

Figura 8.4 Configuración de Nutch mediante cygwin. 97

Figura 8.5 Configuración de Nutch mediante cygwin - Variable de entorno..... 97
Figura 8.6 Establecimiento de variable de entorno desde Windows. 98
Figura 8.7 Establecimiento de las páginas a indexar. 98
Figura 8.8 Configuración de Nutch mediante cygwin (2). 99
Figura 8.9 Configuración del archivo crawl-urlfilter.txt 99
Figura 8.10 Configuración del archivo nutch-site.xml 100
Figura 8.11 Ejecución del crawl desde cygwin. 101
Figura 8.12 Configuración del archivo ‘nutch-site.xml’. 102
Figura 8.13 Interfaz de consulta de Nutch. 102
Figura 8.14 Asistente instalación de Nutch (1). 103
Figura 8.15 Asistente instalación de Nutch (2). 103
Figura 8.16 Símbolo de Regain en la barra de herramientas de Windows. 104
Figura 8.17 Configuración de Regain - Directorios a indexar. 104
Figura 8.18 Configuración de Regain - Páginas a indexar. 105

Índice de Tablas

Tabla 1-1 Análisis de las herramientas candidatas. 25
Tabla 2-1 Evaluación del Sistema de búsqueda para OER. 35
Tabla 2-2 Escala de evaluación para medir la relevancia (Olvera, 2000). 38
Tabla 3-1 Frecuencia de aparición de los criterios (Smith, 1997). 47
Tabla 3-2 Criterios para la selección de los repositorios. 48
Tabla 3-3 Calificación obtenida para los repositorios estudiados 50
Tabla 4-1 Características del equipo. 54
Tabla 4-2 Características del software utilizado. 54
Tabla 5-1 Calificación inicial del SRI. 63
Tabla 5-2 Calificación de los usuarios para el SRI I 64
Tabla 5-3 Calificación de los usuarios para el SRI II 65
Tabla 5-4 Comparación de los resultados obtenidos al evaluar el SRI 66
Tabla 5-5 Número de recursos por categoría y repositorio. 68
Tabla 5-6 Necesidades de información – Preguntas a realizarse. 69
Tabla 5-7 Elaboración de la sintaxis de búsqueda. 70
Tabla 5-8 Matriz - Enlaces Relevantes por pregunta. 72
Tabla 5-9 Documentos relevantes recuperados – SRI NUTCH. 73
Tabla 5-10 Documentos relevantes recuperados – SRI REGAIN. 73
Tabla 5-11 Cálculo de precisión y exhaustividad de NUTCH. 74
Tabla 5-12 Cálculo de la precisión y exhaustividad de REGAIN. 75
Tabla 5-13 Valores de precisión y exhaustividad obtenidos en la evaluación de los sistemas de búsqueda. 75
Tabla 8-1 Categorías de recursos disponibles en repositorio EduTube. 106
Tabla 8-2 Categorías de recursos disponibles en repositorio MIT. 107
Tabla 8-3 Categorías de recursos disponibles en repositorio Merlot 108
Tabla 8-4 Categorías de recursos disponibles en repositorio Connexions. 109
Tabla 8-5 Categoría de antropología en repositorio MIT. 110
Tabla 8-6 Categoría de antropología en repositorio EduTube. 116
Tabla 8-7 Categoría de antropología en repositorio Merlot. 118
Tabla 8-8 Categoría de antropología en repositorio Connexions. 121

1.Estudio de Sistemas de Búsqueda para Recursos Educativos Abiertos

Propósito

1. Estudiar los Recursos Educativos Abiertos, los factores que intervienen en la búsqueda de los mismos y las posibles maneras de recuperarlos a través de la WWW.

2. Conocer la estructura y funcionamiento de un Sistema de Recuperación de Información, el uso de tecnologías semánticas para la recuperación de contenido, y un conjunto de herramientas que faciliten dicho proceso.

Introducción

Los Recursos Educativos Abiertos (OER, Open Educational Resources por sus siglas en inglés) se definen como "materiales y recursos educativos ofrecidos libre y abiertamente para que cualquiera los pueda usar", esta definición fue dada por la UNESCO en el año 2002 (UNESCO, 2002). Aunque por lo general no existe una definición en común ya que han sido muchos los autores que han propuesto diferentes definiciones, este tema se abordará en una sección posterior.

Actualmente, muchas instituciones y universidades cuentan con páginas web, blogs, wikis, etc.; y tanto las instituciones como los usuarios incrementan sus necesidades para poder acceder a los recursos de la web, pero para ello deben contar con las herramientas necesarias para encontrar y organizar la información de una forma descentralizada. La Universidad Técnica Particular de Loja (UTPL) promueve la iniciativa de colocar estos recursos en la web a través del un proceso basado en el uso de software y servicios sociales; el problema es que todos estos recursos se encuentran en diferentes lugares por lo que no es posible realizar una búsqueda global de todos estos recursos al mismo tiempo.

Generalmente los usuarios que realizan búsquedas en internet de materiales educativos, lo suelen hacer de dos formas, ya sea una búsqueda a gran escala (por ejemplo, google o yahoo) o búsquedas especializadas, como por ejemplo sitios de herramientas especificas (ccLearn, 2009); aunque esta segunda forma de buscar es mucho menos utilizada que la primera, resulta mucho más efectiva al momento de buscar OER's, ya que se basa en la búsqueda de datos especiales introducidos por el usuario, esto proporciona un valor adicional a la búsqueda.

La mayoría de los motores de búsqueda populares utilizan un índice de texto y enlaces en las páginas para obtener los resultados. Esto funciona muy bien para la mayoría de las búsquedas de información general. Sin embargo no para la búsqueda de recursos educativos; en este aspecto por ejemplo, los educadores se han interesado en determinados tipos de materiales que tienen ciertos atributos comunes, tales como el tipo de público para el cual fueron diseñados los materiales, la cantidad de tiempo que se tarda en aplicar una lección, etc.

En este capítulo también se mencionan las técnicas de recuperación de información empleadas en internet, las cuales proceden de las empleadas en los sistemas de búsqueda tradicionales; es por esto que han surgido grandes problemas cuando se realizan operaciones de

recuperación con ellos; esto se debe a que el entorno de trabajo no es el mismo y las características de los datos difieren considerablemente; así mismo en la web surgen problemas como en denominado spamming[1] o el enorme tamaño de los índices de los sistemas de búsqueda (Martínez, 2002). Es aquí donde entra en juego la semántica; los datos estructurados mejoran notablemente la recuperación de información; y al tratarse de una web cada vez más descentralizada, el sistema de búsqueda a implementarse realizará sus búsquedas sobre diferentes repositorios.

Este capítulo se divide en 5 secciones, y se estructura como sigue: En la sección 1.1 se hace referencia a los recursos educativos abiertos, con el fin de acercar al lector al problema objeto de estudio; en la sección 1.2 se hace una introducción de los sistemas de búsqueda; en la sección 1.3 se aborda concretamente lo que es la búsqueda de OER, las diferentes maneras de poder realizar las búsquedas y los metadatos que en ellas intervienen; en la sección 1.4, se analizan algunas herramientas de búsqueda, y finalmente, en la sección 1.5 se realiza una comparación de las herramientas estudiadas y una selección de las mismas.

[1] Los constructores de páginas web insertan en la descripción de los mismos términos que no tienen nada que ver con el contenido de las mismas, provocando que los usuarios recuperen estas páginas "truncadas", cuando pretenden recuperar documentos de otra temática.

1.1 Introducción a Open Educational Resources (OER)

1.1.1 Definición de OER

La frase OER (Open Educational Resources) fue adoptada por primera vez en el año 2002 por la UNESCO, la cual define a los OER como: "*Open educational resources* son materiales y recursos educativos ofrecidos libre y abiertamente para que cualquiera los pueda usar" (UNESCO, 2002); y aunque no existe una definición general, una muy comúnmente aceptada es: "Recursos para enseñanza, aprendizaje e investigación que residen en un sitio de dominio público o que se han publicado bajo una licencia de propiedad intelectual que permite a otras personas su uso libre o con propósitos diferentes a los que contempló su autor" (Hewlett, 2005). Estos recursos son de tres tipos: contenidos educativos o también llamados contenidos de aprendizaje, herramientas y recursos de implementación.

En cada uno de los tres tipos de recursos mencionados, los Recursos Educativos Abiertos pueden estar compuestos por:

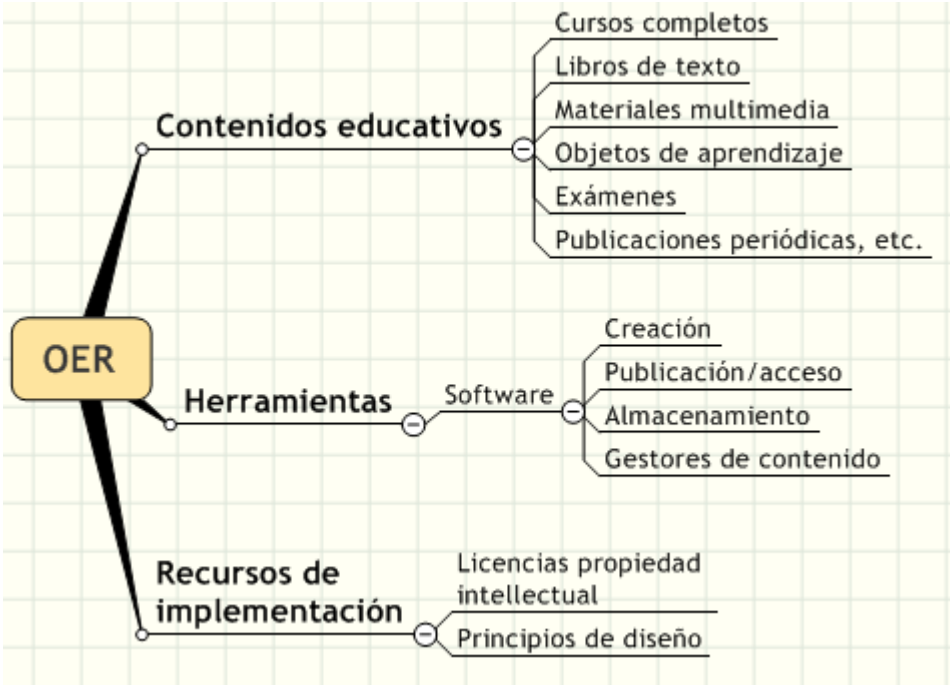

Figura 1.1Tipos de OERs (Hewlett, 2005).

1.1.2 Ventajas y Desventajas de los OER

En (Baker, 2008) se identifican algunas ventajas y desventajas de los OERs, entre las principales ventajas se tienen:

- Promueven iniciativas pedagógicas
- Amplían el uso de alternativas a los libros de texto y a su vez mantienen la calidad educativa.
- Minimiza el costo de los materiales de los cursos para los estudiantes.

Y entre algunas desventajas de los OER están:

- La calidad de los materiales disponibles es inconsistente; es decir no cumplen con todos los estándares o la información que proporcionan no siempre es la correcta.
- No hay una norma común para comprobar la precisión y calidad del OER.
- Necesidad de verificar la precisión del contenido. Antes de publicar cualquier recurso educativo en un repositorio, se debe verificar que no viole derechos de autor y que tenga una licencia de uso libre.
- Requiere adaptación para cumplir con los requerimientos del departamento y/o currículo de la universidad. Antes de publicar cualquier recurso educativo en un repositorio, este debe ser previamente analizado para así determinar si cumple con los estándares y exigencias del repositorio.

1.1.3 Licenciamiento

Todos los OER deben contar con un tipo de licencia de contenido abierto para poder ser publicados o utilizados en los diferentes repositorios y así poder ser compartidos libremente, generalmente suelen ser publicados bajo dos tipos de licencias mencionadas a continuación:

- Licencias Creative Commons (CC)[2]

  Creative Commons ha lanzado varias licencias de derecho de autor conocidas como las licencias Creative Commons. A nivel mundial, una licencia de CC responde a dos preguntas diferentes: ¿Se permite el uso comercial de su trabajo? ¿Permite modificaciones de su trabajo? Para la segunda pregunta, la respuesta puede ser Sí, No o en "Compartir Igual" (que significa que el licenciante permite a otros distribuir trabajos derivados sólo bajo una licencia idéntica a la de la obra original). Este mecanismo también se conoce como copyleft (Bekkers, 2008).

- Licencias GNU

  Las licencias GNU (General Publical License) son utilizadas ampliamente para contenido libre y software de código abierto (FOSS), así como para la documentación. La Licencia pública general GNU (GNU GPL), es utilizada ampliamente como licencia de software libre, fue escrito originalmente por Richard Stallman para el proyecto GNU. La licencia GPL es el más popular y conocido ejemplo del tipo de licencia copyleft fuerte (GNU, 2007).

[2] Creative Commons: <http://creativecommons.org/>

El tema de licenciamiento es de vital importancia al momento de querer buscar OERs; antes de poder usar cualquier contenido que hemos logrado recuperar mediante una búsqueda, se debe comprobar que dicho material utiliza una licencia OER, es decir con la cual podamos reutilizar el material. Es aquí donde el poder buscar materiales en repositorios dedicados a OERs resulta mucho más beneficioso, ya que todos los materiales disponibles en estos repositorios han sido previamente analizados y disponen de una licencia de uso libre.

1.1.4 Repositorios para Recursos Educativos Abiertos

En general, un repositorio es un lugar central en el que se almacenan datos. Puede englobar múltiples bases de datos o ficheros para su distribución en una red, o puede ser un lugar accesible para el usuario sin necesidad de navegar por red (Córcoles, et al., 2007a).

En el contexto educativo, se entiende a un repositorio como un recurso para acceder a OER, donde se encuentran una gran cantidad de contenidos educativos valiosos y reutilizables para una gran variedad de temas tanto para cursos completos como para objetos de aprendizaje más pequeños. Los repositorios soportan mecanismos para importar, exportar, identificar, almacenar y recuperar conocimientos digitales. Los repositorios educativos contienen sobre todo objetos de aprendizaje y materiales educativos o datos para la investigación (Córcoles, et al., 2007a).

Una pregunta interesante en este punto es: ¿Por qué no simplemente almacenar los OERs en cualquier sitio web?; esto se debe principalmente a la visibilidad de los datos en los motores de búsqueda, y aunque los metadatos apropiados se usen para asegurar que los motores de búsqueda lleguen a la página usando las palabras clave adecuadas, la gran cantidad de páginas similares disponibles en Internet hacen que sea muy difícil encontrar un documento específico entre tanta información. Incluso cuando el número de recursos es grande (por ejemplo, centenares), puede resultar único intentar publicarlos en un repositorio temático. Esto asegura un mayor grado de visibilidad y contribuye a mantener los repositorios, mejorándolos en términos cuantitativos y cualitativos. Esto es mucho mejor que publicar los recursos de aprendizaje en un espacio público (páginas web, wikis u otros formatos) y después esperar que los posibles usuarios lo encuentren a través de un motor de búsqueda (Google, por ejemplo) (Córcoles, et al., 2007b).

Para un estudio más detallado sobre los repositorios educativos y de cómo poder buscar y encontrar OERs véase (Córcoles, et al., 2007a).

1.1.5 Estado Actual – Iniciativa de OER

Normalmente los recursos almacenados en los repositorios, son contenidos que ofrecen materiales estáticos en formatos cerrados, es decir, materiales que no pueden ser modificados o materiales que no se pueden copiar. Es por esto que los primeros repositorios que abrieron sus contenidos (openCourseWare) como por ejemplo el MIT OpencourseWare, han ganado mucha reputación, pero esto solo es el primer paso a seguir, ya que no sola basta con tener un repositorio abierto para OER, sino que se tiene que crear, mantener, difundir, actualizar y promover los OER.

Algunas iniciativas ya han surgido con el tema de OER, este es el caso del proyecto OLCOS (Open e-Learning Observatory Services), el cual desarrolla una serie de actividades para promover la creación, así como el compartir y reutilizar los Recursos Educativos Abiertos (OER). OLCOS, es un proyecto de investigación cofinanciado por la Unión Europea en el programa de e-learning. El proyecto OLCOS hasta el momento ha producido como resultado el informe "Open Educational Practices and Resources: OLCOS Roadmap 2012" y los tutoriales en soporte wiki para la adopción de prácticas OER para facilitar la autoformación a creadores y usuarios de contenidos educativos abiertos.

El informe "OLCOS Roadmap 2012", se publicó en enero de 2007 y en él se establece una hoja de ruta para los recursos educativos abiertos. Ofrece un estado del arte actual y visualiza posibles tendencias en el campo de los recursos educativos abiertos, empezando por las definiciones aceptadas para dicho término. El segundo resultado del proyecto; los tutoriales, es la parte más aplicada del proyecto OLCOS. En una plataforma wiki se han creado materiales didácticos para la formación de profesores y estudiantes en la búsqueda, creación, reutilización y generación de licencias para recursos educativos abiertos (Ferran et al., 2009).

Así como OLCOS, muchas instituciones ya han comenzado a promover esta iniciativa, este el caso de Creative Commons con su herramienta de búsqueda de OER llamada, DiscoverEd, que se encuentra en una versión beta, y trabaja sobre repositorios con los cuales mantiene contacto (ccLearn, 2009). En Latino-América también se han realizando proyectos de este tipo, concretamente en Monterrey – México, se ha desarrollado el Knowledge Hub[3], que es una base de conocimiento que provee de un catálogo de colecciones de Recursos Educativos Abiertos disponibles en internet. Este catálogo es creado por la academia del Tecnológico de Monterrey, así como por Universidades e Instituciones educativas afiliadas al proyecto como una propuesta para enriquecer los cursos académicos, mejorar la práctica educativa y apoyar a disminuir la brecha en educación a nivel mundial.

El tema de afiliarse o establecer una alianza con repositorios OER resulta muy llamativo teniendo en cuenta que los educadores y las organizaciones educativas ya gastan tiempo y esfuerzos de evaluación al publicar dichos recursos; ya que este proceso es laborioso y costoso, la primera cuestión en el desarrollo de un sistema de búsqueda especializada es decidir qué recursos incluir; por lo que una herramienta de descubrimiento escalable tendrá que aprovechar la experiencia de la comunidad en general. En el capítulo 4 de esta memoria, se indica el proceso seguido y los criterios aplicados para seleccionar los repositorios de OERs más adecuados para este proyecto.

1.2 Sistemas de Búsqueda

Los sistemas de búsqueda se suelen clasificar principalmente en dos tipos: directorios o índices temáticos y motores o buscadores de contenido, aunque cada vez son más difusos los límites entre unos y otros. Además de índices y motores de búsqueda, existen también los

[3] Knowledge Hub (http://khub.itesm.mx/)

denominados metabuscadores, buscadores en paralelo, megamotores o metaservidores de información en Internet. Estos sistemas van más allá de los buscadores: admiten una consulta y se encargan de lanzarla a diferentes sistemas de búsquedas públicos que hay en Internet.

Los metabuscadores ofrecen detalles de las respuestas de cada uno de los servicios, o bien el listado completo de coincidencias que constituyen (al menos en teoría) las mejores respuestas a la pregunta formulada. Generalmente no se obtiene toda la potencia de cada uno de ellos (dado que los formatos de consulta varían) pero pueden ser útiles cuando no se han tenido suerte en la búsqueda en otros servicios, para buscar por una materia poco común, o para realizar búsquedas exhaustivas (Hernández & Méndez, 2009).

El proceso llevado a cabo por cualquier sistema de búsqueda se puede resumir en las siguientes fases: (Lamarca, 2009)

- recogida y análisis de datos (indización y/o clasificación por categorías)
- búsqueda propiamente dicha
- recuperación

Tanto la recogida de datos como el análisis de los mismos pueden hacerse bien de forma manual, o bien de forma automática. Para la recogida de datos manual, los Índices suelen presentar un cuestionario en línea para que la persona u organización que quiera darse de alta identifique y clasifique su página web.

Los motores de búsqueda suelen utilizar la recogida de datos automática rastreando la red, otros piden la dirección URL para darse de alta. Disponen de un robot que visita y analiza la página principal y todas las páginas enlazadas y que suele ser capaz de leer las etiquetas META o metadatos y extraer toda la información contenida en ellas mediante el lenguaje HTML. Sin embargo, muchas páginas no disponen de tales etiquetas. Con dicha información, el buscador es capaz de indizar palabras clave como el título, idioma, autor, propietario, localización, temas, etc.

Existen sistemas de búsqueda que mezclan estas dos funciones y ofrecen búsquedas por medio de un índice temático y búsquedas libres por palabras clave. Un buen sistema de búsqueda debe permitir flexibilidad en las búsquedas ofreciendo la posibilidad de elegir entre búsquedas mediante clasificación temática o por medio de formularios. Los formularios deben ofrecer tanto búsquedas sencillas como búsquedas más complejas que permitan algún tipo de herramientas como truncado de palabras, operadores booleanos, términos compuestos, acotación de búsquedas, etc. y con diferentes campos de búsqueda en los que se requiera un lenguaje libre o controlado (título, palabras clave, idioma, localización, tipo de información, etc.). También deben ser capaces de controlar el vocabulario para deshacer ambigüedades, sinonimias, polisemias[4], etc. Además, los sistemas de búsqueda, deben presentar los resultados de la búsqueda de una forma también flexible permitiendo varios criterios de

[4] En lingüística se presenta cuando una misma palabra o signo lingüístico tiene varias acepciones.

aparición y ordenación de los datos y ofreciendo diferentes formatos para que el usuario elija el que se ajusta a su gusto y necesidades (Lamarca, 2009).

1.2.1 Definiciones

Recuperación de Información

"La recuperación de información (IR), por sus siglas en inglés (Information Retrieval), es la ciencia de la búsqueda de información en documentos, búsqueda de los mismos documentos, la búsqueda de metadatos que describan documentos, o, también, la búsqueda en bases de datos, ya sea a través de internet, intranet, para textos, imágenes, sonido o datos de otras características, de manera pertinente y relevante" (Baeza-Yates & Ribeiro-Neto, 1999). Otra definición de Recuperación de Información, "es la disciplina que tiene por objetivo el desarrollo de sistemas que almacenen grandes cantidades de documentos de tal forma de permitir una eficiente recuperación de aquellos documentos relevantes a las necesidades de información de sus usuarios" (Cursada, 2009).

Sistemas de Recuperación de Información

Los Sistemas de Recuperación de Información (SRI) son una clase de sistemas de información que tratan con bases de datos compuestas por documentos y procesan las consultas de los usuarios permitiéndoles acceder a la información relevante en un intervalo de tiempo apropiado. Estas consultas son sentencias formales mediante las cuales el usuario expresa sus necesidades de información, formuladas usando un lenguaje de consulta. Un SRI está compuesto por tres componentes principales: la base de datos documental, el subsistema de consultas y el mecanismo de emparejamiento o evaluación (López, 2006).

1.2.2 Proceso de Recuperación de Información

Un proceso de recuperación de información comienza cuando un usuario introduce una consulta en el sistema. Las consultas son declaraciones oficiales de las necesidades de información; por ejemplo, las cadenas de búsqueda en los motores de búsqueda web. Las consultas del usuario se comparan con la información de la base de datos. Dependiendo de la aplicación, los objetos de datos pueden ser, por ejemplo, documentos de texto, imágenes o videos. A menudo, los mismos documentos no se conservan o almacenan directamente en el sistema de recuperación, sino que son representados en sistemas sustitutos de documentos o metadatos (Byers & Kormann, 2009).

La mayoría de los sistemas de recuperación de información calculan una puntuación numérica de lo bien que cada objeto coincide con la consulta en la base de datos, y clasifican dichos objetos de acuerdo a este valor. Los objetos mejor clasificados serán mostrados al usuario, pudiendo entonces el proceso ser repetido si el usuario desea refinar la consulta (Frakes & Baeza-Yates, 1992).

El concepto de recuperación de información es bastante simple, tal como se muestra en la **Figura 1.2.** Sin embargo, los procesos involucrados en la determinación de la relevancia de

una consulta suelen ser complejos, especialmente cuando se está tratando con información que carece de una estructura definida.

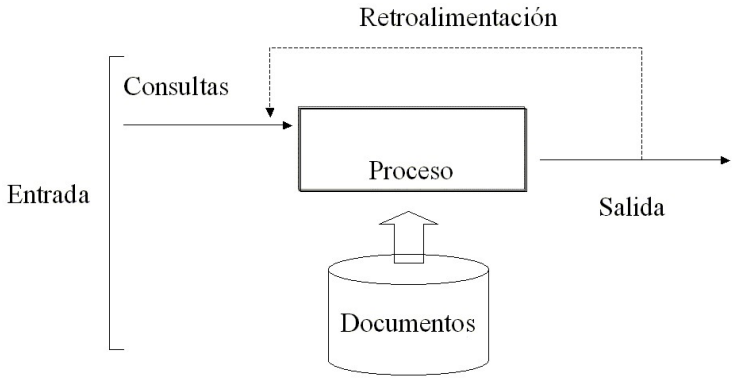

Figura 1.2 Proceso de Recuperación de Información (Salazar & Pinto, 2003).

1.2.3 Elementos Clave de las Búsquedas

Es necesario tener en cuenta los elementos clave que permiten hacer la búsqueda, determinando un mayor grado de pertinencia y precisión, como son: los índices, palabras clave, tesauros y los fenómenos que se pueden dar en el proceso como son el ruido y silencio documental. Uno de los problemas que surge en la búsqueda de información es si lo que un usuario recupera es "mucho o poco" es decir, dependiendo del tipo de búsqueda se pueden recuperar multitud de documentos o simplemente un número muy reducido. A este fenómeno se denomina Silencio o Ruido documental (Pinto, 2009).

- **Silencio documental:** Son aquellos documentos almacenados en la base de datos pero que no han sido recuperados, debido a que la estrategia de búsqueda ha sido demasiado específica o que las palabras clave utilizadas no son las adecuadas para definir la búsqueda.
- **Ruido documental:** Son aquellos documentos recuperados por el sistema pero que no son relevantes. Esto suele ocurrir cuando la estrategia de búsqueda se ha definido demasiado genérica.

A continuación se describe brevemente lo que son índices, palabras clave y tesauros; para una explicación más detallada sobre estos términos, véase (Pinto, 2009).

- **Índices**
  Listado de términos normalizados que representan el contenido de un recurso. Algunos tipos son:
- **Palabras clave (Keywords)**
  Término significativo en lenguaje natural que representa el contenido del documento. En la búsqueda de información esta opción es esencial ya que permite acotar y precisar información. El problema recae en definir la palabra exacta que representa el contenido, por ello es conveniente utilizar especificadores.

- **Tesauros**

Es un listado terminológico controlado sobre un área o ámbito de conocimiento que mantiene entre sí relaciones semánticas y genéricas. Su principal característica es que los términos están ordenados jerárquicamente, permitiendo la precisión terminológica en la búsqueda de información.

1.2.4 Técnicas de Recuperación de Información

La naturaleza de las técnicas que se pueden utilizar para recuperar información, provienen de la estadística, aprendizaje automático e inteligencia artificial; en la **Figura 1.3** se identifican las más conocidas.

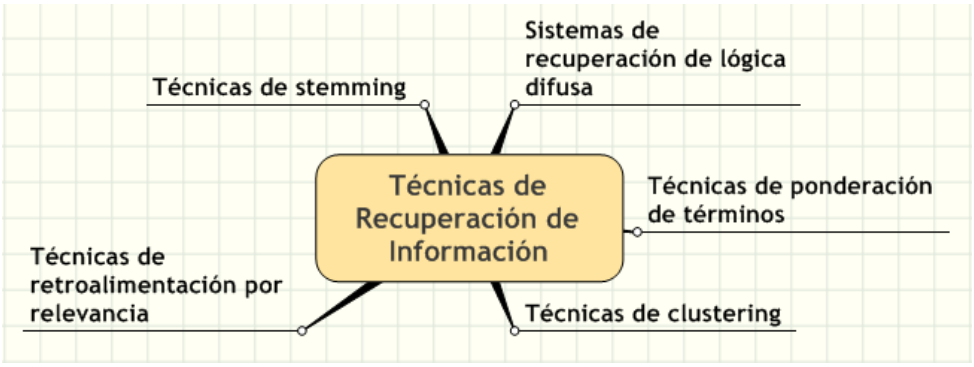

Figura 1.3 Técnicas de recuperación de información (Pinto, 2009).

- **Sistemas de recuperación de lógica difusa**

Esta técnica permite establecer consultas con frases normales, de forma que la máquina al realizar la búsqueda elimina signos de puntuación, artículos, conjunciones, plurales, tiempos verbales, palabras comunes (que suelen aparecer en todos los documentos), dejando sólo aquellas palabras que el sistema considera relevantes.

- **Técnicas de ponderación de términos**

Es común que unos criterios en la búsqueda tengan más valor que otros, por tanto la ponderación pretende darle un valor adecuado a la búsqueda dependiendo de los intereses del usuario. Los documentos recuperados se encuentran en función del valor obtenido en la ponderación. El valor depende de los términos pertinentes que contenga el documento y la frecuencia con que se repita. De forma que, el documento más pertinente de búsqueda sería aquel que tenga representado todos los términos de búsqueda y además el que más valor tenga repetidos más veces, independientemente de donde se localice en el documento.

- **Técnica de clustering**

Es un modelo probabilístico que permite las frecuencias de los términos de búsqueda en los documentos recuperados. Se atribuyen unos valores (pesos) que actúan como agentes para agrupar los documentos por orden de importancia, mediante algoritmos ranking.

- **Técnicas de retroalimentación por relevancia**

Esta técnica pretende obtener el mayor número de documentos relevantes tras establecer varias estrategias de búsqueda. La idea es que, tras determinar unos criterios de búsqueda y observar los documentos recuperados se vuelva a repetir nuevamente la consulta pero esta vez con los elementos interesantes, seleccionados de los documentos primeramente recuperados; en las técnicas de retroalimentación por relevancia se usa el llamado algoritmo genético.

- **Técnicas de stemming**

Morfológicamente las palabras están estructuradas en prefijos, sufijos y la raíz. La técnica de Stemming lo que pretende es eliminar las posibles confusiones semánticas que se puedan dar en la búsqueda de un concepto, para ello trunca la palabra y busca solo por la raíz.

1.2.5 Modelos de Recuperación de Información

Los principales modelos clásicos de recuperación de información son: modelo Booleano, modelo Espacio Vectorial, modelo Probabilístico y modelo Booleano extendido o modelo Difuso, a continuación se los describe brevemente, para un análisis más detallado véase (López, 2006).

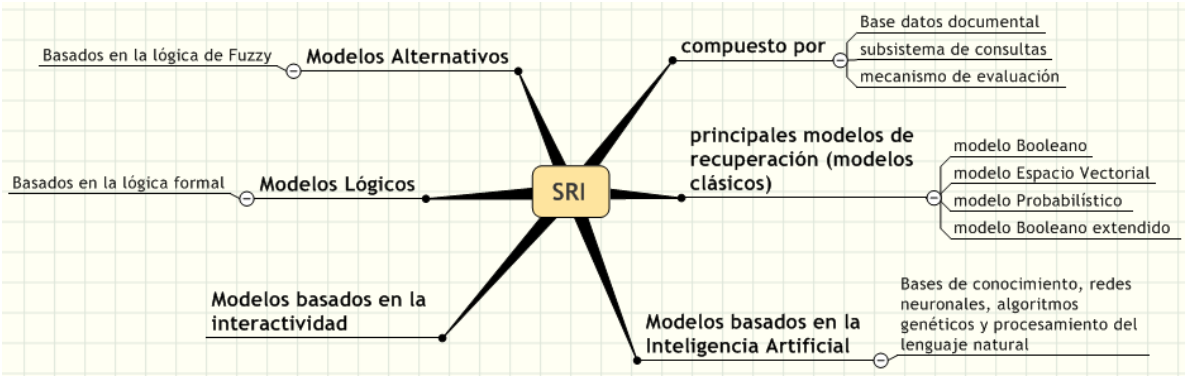

Figura 1.4 Clasificación de los SRI (López, 2006) (Martínez F. J., 2002).

1.2.6 Arquitectura General de un Sistema de Recuperación

En (Mel'nikov et al., 2008) se menciona que un motor de recuperación perfecto deberá satisfacer los siguientes requisitos fundamentales:

- Usabilidad
- El índice deberá estar correctamente organizado y estructurado
- Búsqueda y reacción rápida en la base de datos
- Fiabilidad y exactitud en los resultados de búsqueda

La arquitectura general de un moderno Sistema de Recuperación de Información (SRI) de Internet puede ser presentado de forma esquemática como sigue:

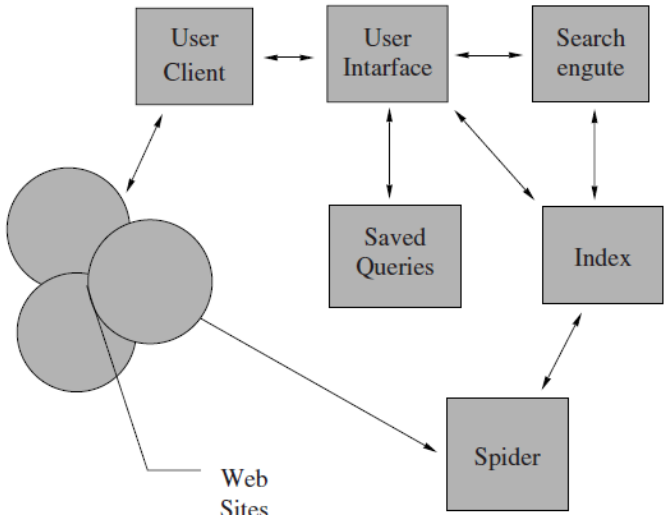

Figura 1.5 Arquitectura general de un SRI (Mel'nikov et al., 2008).

En el esquema mostrado se indican los siguientes componentes de un SRI, en (Mel'nikov et al., 2008) se estudia a mayor detalle dicho esquema:

Un **usuario cliente** es un programa que examina la información de determinados recursos. Este programa ofrece navegación en los documentos, archivos FTP, listas de correo, grupos de noticias, etc. Todos estos recursos de información son, a su vez, los objetos de la búsqueda de un SRI.

La **Interfaz de usuario**. La interfaz de usuario no sólo representa el programa de navegación, sino que en el caso de un sistema de recuperación de información es también el método de comunicación del usuario con el mecanismo de recuperación del sistema, es decir, el sistema que forma las consultas y examina los resultados de búsqueda (Mel'nikov et al., 2008). La interfaz de usuario se divide en dos partes (Ramírez, 2009):

- Interfaz de consulta.
- Interfaz de respuesta.

La interfaz de consulta básica es generalmente una caja de texto donde el usuario puede escribir una o varias palabras. La interfaz de respuesta tiene por objetivo mostrar las páginas encontradas más relevantes para el usuario, esta interfaz deberá mostrar elementos como son el URL, tamaño, fecha de indexación, título y las primeras líneas de la página o documento encontrado (Ramírez, 2009).

Motor de búsqueda. Un motor de búsqueda se utiliza para: (1) la traducción de la consulta introducida por el usuario, la cual se prepara en un idioma de consulta formal para que sea entendida por el sistema de recuperación de información, (2) la búsqueda de enlaces a recursos de información de Internet, (3) la entrega de resultados de búsqueda para el usuario (Mel'nikov et al., 2008). La misión del motor de búsqueda es la de analizar la consulta del usuario, buscar en el índice las páginas relacionadas a la búsqueda y

de ordenar según la relevancia estimada, criterios de localización, frecuencia de aparición y popularidad de las páginas (Ramírez, 2009).

Indexador. Es el encargado de obtener la representación interna de cada página encontrada por el crawler. Hasta que una página no ha sido indexada no está disponible para ser devuelta como resultado de una búsqueda (Ramírez, 2009).

Consultas (Queries). Las consultas se guardan en la base de datos personal del usuario. La tramitación de cada consulta tarda un cierto tiempo, por lo tanto, es muy importante guardar las consultas para que el sistema responda bien.

Robot Indexador (spider). Se utiliza para explorar el Internet. Este programa es la principal fuente de información sobre el estado de los recursos de información de la red. La presencia de "robots de búsqueda" es la principal diferencia en los sistemas de recuperación de los directorios de búsqueda. El más conocido directorio de búsqueda es Yahoo.com (Mel'nikov et al., 2008).

El robot indexador también es llamado crawler. Se ejecuta en una máquina local y envía peticiones a los servidores Web, realizando visitas periódicas; comienza con un conjunto de URLs, ya sea por ejemplo haciendo un recorrido en profundidad en anchura. El crawler permite que se le proporcionen direcciones de sitios Web. Los buscadores combinan las técnicas anteriores con medidas de popularidad para decidir el orden en que se visitan las páginas. El objetivo es recorrer las páginas de mayor calidad; la popularidad se mide como el número de enlaces que apuntan a la página. La frecuencia entre visitas es variable, pudiendo ser desde días a meses (Ramírez, 2009).

Estos componentes corresponden a una arquitectura centralizada: i) el crawler (robot, spider, entre otros) corren localmente en la máquina de búsqueda, recorriendo la Web mediante pedidos a los servers y trayendo el texto de las páginas Web que va encontrando ii) El indexador corre localmente y mantiene un índice sobre las páginas que le trae el crawler iii) El motor de búsqueda también corre localmente y realiza las búsquedas en el índice, retornando básicamente los URLs rankeados iv) La interfaz corre en el cliente (cualquier parte de la Web) y se encarga de recibir la consulta, mostrar los resultados, retroalimentación, etc. (Ramírez, 2009).

1.2.7 Búsqueda Sobre Datos Estructurados

En muchos de los sistemas de búsqueda actuales, resulta demasiado laborioso el poder gestionarlos y ampliarlos más allá de un solo sitio o conjunto de recursos, esto se debe a que los datos en la mayoría de los sitios no están debidamente estructurados o simplemente no lo están; sin embargo para poder mejorar la búsqueda de recursos digitales, y particularmente la búsqueda de OER's, es necesaria de alguna forma una estructuración de los datos si los resultados de búsqueda de materiales educativos deben ser mejorados (ccLearn, 2009). Por ejemplo, los criterios de diseño y los parámetros de evaluación de un recurso educativo son atributos fundamentales para los recursos educativos, y estos requieren actualmente del etiquetado humano. Así, uno de los retos al diseñar una herramienta de búsqueda, es que esta

aproveche los datos disponibles estructurados (también llamados metadatos). Esta información debe ser archivada de una manera que pueda ser descubierta y aprovechada fácilmente.

El obstáculo para las personas que buscan recursos educativos en internet no es la falta de materiales, sino la dificultad para descubrir esos materiales; usualmente las herramientas que utilizamos para descubrir estos recursos son los motores de búsqueda.

Aunque la mayor parte de los motores de búsqueda web a gran escala se basan en algoritmos computacionales para extraer el contenido escrito de un texto, las aplicaciones de búsqueda más específicas pueden tomar mayor ventaja de datos estructurados para proporcionar mejoras más flexibles y búsquedas específicas y concretas, también conocidas como búsquedas mejoradas. El problema aquí, desafortunadamente es que la creación manual de información estructurada sobre todos los recursos educativos disponibles sería demasiado costosa y muy prolongada. Este hecho significa que muchos recursos potencialmente valiosos están asociados con los datos estructurados en una mínima parte o casi nada. Por otra parte, los datos estructurados que existen pueden ser muy variables en su contenido y en sus especificaciones; y aunque esto representa un verdadero desafío, los datos estructurados pueden mejorar la exhibición de los materiales apropiados y deseados, por lo tanto, una herramienta de búsqueda que realice múltiples consultas en distintos repositorios de recursos educacionales de alguna manera debe capitalizar los datos disponibles sin dejar de ser robusta (ccLearn, 2009).

En general, un sistema de búsqueda de material educativo sobre datos estructurados deberá soportar algunos de los lenguajes descritos a continuación:

RDF

RDF (Resource Description Framework), es un modelo estándar para el intercambio de datos en la Web, tiene características que facilitan la fusión de los datos incluso si los esquemas subyacentes son diferentes. El uso de este modelo simple, permite que los datos estructurados y semi-estructurados puedan ser combinados, expuestos y compartidos a través de diferentes aplicaciones (RDFWorkingGroup, 2004).

RDF está especialmente destinado a la representación de metadatos sobre los recursos Web, tales como el título, el autor, la fecha de modificación de una página web, los derechos de autor. RDF está diseñado para situaciones en las cuales esta información deba ser procesada, en lugar de solo ser mostrada a la gente (W3C, 2004).

RDFa

RDFa es una especificación de atributos para expresar datos estructurados en cualquier lenguaje de marcado. RDFa es en esencia una generalización de los atributos de los elementos *meta* y *link* de XHTML; siendo por esto la letra minúscula que se añade a las siglas RDF, que hace referencia a la inicial la palabra “atributos”. RDFa es desarrollado y propuesto por el W3C (World Wide Web Consortium) (W3C, Sobre el W3C, 2010). El objetivo de RDFa es hacer que cualquier estructura RDF sea representable en el más puro XHTML. RDFA requiere

XHTML, y por ende una nueva forma de URI[5] (***U**niform **R**esource **I**dentifier*), llamada CURIER[6] (Graf, 2007).

Microformatos

En (González, 2008) se dice que los microformatos son "pequeños trozos de HTML basados en formatos de código abierto que permiten la publicación de información de alta fidelidad en la Web; ellos son el modo más rápido y simple de soportar canales de RSS (Really Simple Syndication) y APIs (Application Programming Interface) en un sitio". Además de ayudar a contribuir semántica en el HTML, los microformatos se limitan a introducir pequeños (de ahí su prefijo 'micro') fragmentos de código reutilizable, código que aporta al usuario información válida como son contactos, eventos, etiquetas, etc. de una manera sencilla. El objetivo de los microformatos es hacer útiles los metadatos, tanto a los usuarios como a los agentes de usuarios, buscadores u otros. Cada autor es libre de crear sus propios microformatos, aunque existen algunos que se han generalizado, para revisar a más detalle este punto véase (González, 2008).

RSS

Las últimas versiones de RSS corresponden a las siglas de:

RDF Site Summary (RSS 1.0)

RSS es una ligera descripción de metadatos y un formato para sindicar[7] el contenido en la Web. RSS es una aplicación XML. Un resumen RSS, es como mínimo un documento que describe un 'canal' que consta de elementos URL recuperables. Cada elemento consta de un título, un enlace y una breve descripción (RSS-DEV Working Group). RSS 1.0 es un vocabulario RDF que proporciona una forma liviana, pero potente de describir la información para su distribución oportuna y reutilización a gran escala; RSS 1.0 es quizás la aplicación más amplia de RDF en la web (W3C, 2004).

Really Simple Syndication (RSS 2.0)

RSS 2.0 es un formato de sindicación de contenido web, RSS es un dialecto de XML, todos los archivos de RSS se ajustan a la especificación de XML 1.0 así como a las publicaciones de la página web del W3C. Un canal RSS 2.0 debe tener como mínimo los siguientes elementos (RSS Advisory Board, 2009): Title, Link, Description.

[5] Un URI es una cadena corta de caracteres que identifica inequívocamente un recurso

[6] contracción para *Compact URI*

[7] Toma de datos disponibles en línea para su recuperación y posterior transmisión, agregación o publicación en línea.

Atom

El formato Atom fue desarrollado como una alternativa a RSS. Atom es un formato de documento basado en XML que describe listas de información relacionada conocidas como 'feeds'. Los feeds se componen de una serie de elementos conocidos como 'entradas', cada una con un conjunto ampliable de metadatos. El uso principal de atom es la sindicación de los contenidos web como contenido personal, weblogs, titulares de noticias así como los agentes de usuario (Nottingham & Sayre, 2005). Atom es una manera simple de leer y escribir información en la web, esto permite seguir fácilmente la pista de más sitio en internet en un menor tiempo. Atom está diseñado para ser un estándar universal para la publicación de contenido personal y weblogs (Atom-Enabled Alliance, 2004).

1.3 Búsqueda de OER

Al momento de buscar OER, se presentan 3 posibilidades de búsqueda: (i) desde los motores de búsqueda tradicionales como yahoo o google; (ii) buscadores especializados, los cuales buscan solo en repositorios OER y (iii) según el tipo de material educativo, por ejemplo material audiovisual. Por lo general los usuarios comunes no tienen conocimiento de dichas herramientas de búsqueda especializada, y usualmente utilizan los motores de búsqueda tradicionales; en esta sección se analizarán todas estas alternativas.

Pero antes de usar el contenido que se ha encontrado, se debe comprobar si éste utiliza una licencia y cómo se puede reutilizar el material. A menudo será difícil encontrar información clara respecto al copyright. Si se han encontrado múltiples recursos sin información clara respecto al copyright, es probable que no se pueda reutilizar o modificar dichos recursos. Sin embargo, comprobar la licencia y restricciones de reutilización de cada recurso o documento que se ha encontrado lleva mucho tiempo; es por esta razón que resulta lógico usar herramientas de búsqueda que identifiquen contenidos OER con licencia en cualquier lugar de la Web (Córcoles, et al., 2007a).

1.3.1 Uso de Buscadores Tradicionales

Algunos buscadores que se usan tradicionalmente, permiten también buscar contenidos abiertos; éstos incluyen opciones avanzadas para la búsqueda de OER, es decir recursos que estén disponibles bajo una licencia de Creative Commons. Para ello, por lo general cuentan con una opción de "derechos de uso" en "búsqueda avanzada" o en "preferencias"; esto con el fin de evitar comprobar el copyright y la reutilización de cada recurso que se haya encontrado; se puede especificar de antemano qué tipo de recursos son los que se están buscando (Córcoles, et al., 2007a). Dos de los buscadores más comunes, los cuales posen este tipo de opciones son:

- El buscador Yahoo CC[8]

[8] http://search.yahoo.com/cc/

- y el buscador Google[9]

  En búsqueda avanzada selecciona la opción de derechos de uso que mejor se ajuste a los requerimientos de cada usuario.

1.3.2 Uso de Buscadores Especializados

Al buscar OER, existen algunos campos en concreto que son de interés para la realización de una búsqueda, algunos posibles criterios/guías para seleccionar materiales abiertos son los siguientes (Baker, 2008):

- Calidad del contenido, mérito literario y formato

  Es decir, la importancia educativa de los contenidos.
- Mantiene su relevancia a pesar de que pase el tiempo
- Recibe reseñas favorables
- Autoridad
- Profundidad
- Formatos disponibles: impreso, CD-ROM, en línea, entre otros.
- Nivel de lectura
- Accesibilidad

Actualmente existen diversos buscadores de OER, por nombrar algunos de estos se tiene:

- DiscoverEd Beta (Creative Commons)[10]
- Connexions[11]
- MERLOT[12]
- OERCOMMONS[13]

1.3.3 Uso de Buscadores por Tipo de Recurso

Si se necesitan materiales de tipo audiovisual (imágenes, animación, audio, video) también existen recursos específicos en la Web donde se pueden buscar este tipo de contenidos abiertos. Esto puede ser muy útil si se está preparando nuevos objetos de aprendizaje o unidades didácticas y no se quiere reutilizar objetos enteros. La siguiente lista ofrece material con licencia de Creative Commons para dichos recursos (Córcoles, et al., 2007a):

- Flickr Creative Commons Search[14]
- En el repositorio de imágenes Wikimedia Commons[15]

[9] http://www.google.com.ec/advanced_search?hl=es

[10] http://discovered.creativecommons.org/search/search.jsp?query=&hitsPerPage=10&lang=es

[11] http://cnx.org/content/browse_content/

[12] http://www.merlot.org/merlot/materials.htm

[13] http://www.oercommons.org/search

[14] http://www.flickr.com/creativecommons/

- El Freesoundproject[16] (solo sonidos, no canciones)
- El sitio CCmixter[17] (canciones)

1.3.4 Estándares de Metadatos de Material Educativo

Hoy en día gran parte de los buscadores o recuperadores de información no tienen en cuenta las meta-etiquetas de las páginas, ya que dan mayor peso al contenido de esta que a la descripción que ofrecen los metadatos. El mayor problema que surge con el uso de los metadatos es que estos no se gestionan de una manera automática, es decir, es el propio creador del sitio web el que se tiene que preocupar de las meta-etiquetas (crearlas y gestionarlas), lo cual es un gran problema si los contenidos cambian constantemente, o si estos son muy grandes, y por lo general esto no se realiza (Cid, 2005).

Los metadatos permiten al usuario de los objetos ubicar la información del recurso y recuperarlos de la base de datos de donde se encuentra, algunos metadatos usuales de los objetos de aprendizaje pueden ser la temática, contenidos con los que trabaja, tecnologías de información en los que se apoya (ontología), tipo de actividad que promueve, autor, institución, etc. Los estándares (ISO, ANSI, IEEE) tratan de elaborar abstracciones de alto nivel o arquitecturas que representen toda una gama diversa de implementaciones prácticas de las mismas y utilizando arquitectura de un sistema y especificación (AICC, ADL, IMS, ARIADNE) (Toscano, 2009). Entre las iniciativas que tratan de estandarizar metadatos que permitan describir y recuperar recursos educativos, están (Suárez, 2004):

IEEE/LTSC LOM e IMS Global Learning Meta-Data.

Se trata del estándar por excelencia para Metadatos de Objetos Educativos, y está patrocinado por el Comité de Estandarización de Tecnologías Educativas del IEEE. Alrededor de este estándar es habitual encontrarse con confusiones sobre el nombre y sus autores.

Dublin Core Metadata Initiative.

DCMI[18] (Dublin Core Metadata Initiative) es una organización internacional y virtual albergada por OCLC (Dublin Ohio, EE.UU) que se financia a través de proyectos y subvenciones y en la que el trabajo se realiza por voluntarios de todo el mundo.

Los metadatos son datos acerca de los datos, específicamente el término se refiere a los datos que se utilizan para identificar, describir o localizar recursos de información, sean estos recursos físicos o electrónicos. Dublin Core (DC) es un conjunto de 'elementos' (propiedades) para describir documentos (y por lo tanto para el registro de metadatos). El objetivo de DC es proporcionar un conjunto mínimo de elementos descriptivos que faciliten la descripción y la

[15] http://commons.wikimedia.org/wiki/Main_Page

[16] http://freesound.iua.upf.edu/

[17] http://ccmixter.org/

[18] DCMI. Disponible en: http://dublincore.org/

indexación automatizada de objetos en red, de manera similar a un catálogo de fichas de biblioteca. El conjunto de metadatos DC está destinado para el uso de herramientas de internet como los 'web crawlers' empleados por los motores de búsqueda de la www (W3C, 2004).

Además de Estándares Abiertos para software también hay especificaciones y estándares abiertos para la descripción de contenido educativo, estos se analizan con mayor detalle en (Córcoles, et al., 2007b). Estos estándares se incluyen en muchos de los sistemas y aplicaciones de gestión de contenidos o de aprendizaje y en algunos repositorios de contenido, pero no en la mayoría de aplicaciones wiki.

1.4 Herramientas de Búsqueda

A continuación, se describen algunas de las herramientas para la búsqueda de información y que pueden ser adoptadas para encontrar material educativo.

1.4.1 Courseware Watchdog

Courseware Watchdog[19] es una herramienta basada en ontologías, diseñada para aprovechar los recursos disponibles en la web; es parte del proyecto PADLR (acceso personalizado a Distributed Learning Repositories). Watchdog se basa en dos componentes de recuperación (CRAWLER y EDUTELLA) que permiten al usuario encontrar material de acuerdo a sus intereses. En ambos casos, el componente de navegación de la ontología se utiliza para definir los elementos que deben ser buscados. Mientras que CRAWLER se centra en buscar estas preferencias de los usuarios al buscar los términos pertinentes en todo el texto de los documentos web. La red EDUTELLA se utiliza para intercambiar metadatos sobre los recursos de aprendizaje que se han anotado semánticamente. Courseware Watchdog incluye un rastreador web para recuperar material de aprendizaje de la WWW; los metadatos del proceso de rastreo se almacenan en una ontología. (Tane et al., 2004).

1.4.2 Knowledge Hub

El Knowledge Hub[20] (KHub) es un Nodo Público Multilingüe que indexa y cataloga OERs existentes en el Internet, y que provienen de sitios académicos y de instituciones de reconocimiento internacional. La idea central del KHub es tener una base de datos de recursos educativos abiertos y objetos de aprendizaje disponibles en la red (tales como: presentaciones en PPT, podcast, videos-en-demanda, weblogs, blogs, software, etc.) para asistir en el proceso instruccional y de aprendizaje a nivel mundial (Mortera & Escamilla, 2008). KHub utiliza un conjunto de metadatos definidos por expertos bibliotecarios e informáticos, también cuenta con una serie de herramientas que permiten la construcción de redes sociales para compartir comentarios y dar rangos y puntaje a estos recursos educativos abiertos disponibles en el KHub.

[19] http://cwatchdog.sourceforge.net/

[20] http://www.temoa.info/es/

1.4.3 DiscoverEd

DiscoverEd[21] es un proyecto experimental de ccLearn que ofrece la búsqueda y descubrimiento de recursos educativos en la web. El conjunto de resultados que se obtiene incluye los metadatos, el tipo de licencia de los recursos y toda la información disponible sobre el tema. Lo interesante de este proyecto es que está enfocado a OERs y colabora con otros proyectos para mejorar las capacidades de búsqueda y de descubrimiento de este tipo de recursos. DiscoverED es un prototipo destinado a explorar cómo los datos estructurados pueden ser utilizados para mejorar la experiencia de búsqueda. Los resultados obtenidos en la búsqueda provienen de repositorios institucionales con los que se mantiene contacto, estos repositorios son: i) OER Commons ii) Connexions iii) the Open Courseware Consortium (OCWC) iv) the National Science Digital Library (NSDL) (CreativeCommons, 2009).

1.4.4 Lucene

Lucene[22] es un API de código abierto para la recuperación de información. Está apoyado por la ASF (Apache Software Foundation) y se distribuye bajo la licencia ASL (Apache Software License). Lucene es una librería de búsqueda para Java que permite añadir búsqueda a cualquier aplicación. Esta herramienta es útil para cualquier aplicación que requiera indexado y búsqueda de texto completo. Ha sido ampliamente utilizada por su utilidad en la implementación de motores de búsqueda; por ello, a veces se confunde Lucene con un motor de búsqueda con funciones de "crawling" y análisis de documentos en HTML incorporadas (Hatcher et al., 2008).

Lucene indexa el contenido estructurado (metadatos) de cada recurso, de manera que el recorrido y búsqueda de información se realiza de una forma rápida y sencilla. Lucene soporta la indexación de documentos con formatos: txt, pdf, doc, ppt, rtf, xml y html. Lucene ofrece una API flexible, a través de la cual se añaden, con esfuerzos de programación, capacidades de indización y búsqueda a cualquier sistema que se esté desarrollando (Ramos & Hernández, 2008).

1.4.5 SOLR

SOLR[23] es una plataforma de búsqueda basado en HTTP, de código abierto y parte del proyecto de Apache Lucene. Entre sus características principales se incluyen: búsquedas de texto completo, resaltado de las palabras importantes, búsqueda en facetas, agrupación dinámica (dynamic clustering), integración con bases de datos, y la manipulación de documentos (por ejemplo, Word, PDF). SOLR es escalable, proporciona servicios de búsqueda distribuida y replicación del índice (Lucene, 2010).

[21] http://discovered.creativecommons.org/search/

[22] http://lucene.apache.org/

[23] http://lucene.apache.org/solr/

SOLR está escrito en Java y funciona como un servidor completamente independiente de búsqueda de texto dentro de un contenedor de servlets como por ejemplo Apache Tomcat. SOLR usa la biblioteca de búsqueda 'Lucene Java' como su centro para la indización de texto completo y búsqueda, utiliza REST y APIs en HTTP/XML y JSON que facilitan la utilización de prácticamente cualquier idioma de programación. Finalmente se puede destacar que SOLR tiene una arquitectura basada en plug-ins para cuando se requiere personalización más avanzada (Lucene, 2010).

1.4.6 SIREn

SIREn (Semantic Information Retrieval Engine) es un plugin de Lucene que indexa y consulta microformatos y RDF, así como cualquier documento de texto descrito con diferentes metadatos (Delbru, 2009).

SIREn utiliza la biblioteca de Lucene y SOLR, lo que significa que usando SIREn se pueden aprovechar las características de ambos. SIREN está disponible como software de código abierto bajo la licencia Apache 2.0, la cual permite usar SIREn en los programas tanto comerciales como de código abierto (Delbru, SIREn Presentation, 2009).

1.4.7 Regain

Regain[24] es un motor de búsqueda basado en Lucene, similar a los motores de búsqueda web como Google, con la diferencia de que no busca en toda la web, sino en sus propios archivos y documentos; el crawler de Regain recupera archivos o páginas web, extrae todo el texto y lo coloca en un índice de búsqueda inteligente; todo este proceso es transparente para el usuario; la búsqueda de escritorio viene con su propio 'servidor web'.

Regain está escrito en el lenguaje JAVA, y por lo tanto es aplicable a todas las plataformas compatibles con JAVA. El servidor de búsqueda funciona con JSP (Java Server Pages) y una biblioteca de etiquetas. Regain es liberado bajo licencia LGPL (Licencia pública general); es decir, puede ser utilizado de forma gratuita sin ningún límite temporal, se concede la autorización de usar, desarrollar y personalizar regain, con la única condición de que se proporcione el código fuente libremente para todo el mundo (Schneider & Tesche, 2010).

1.4.8 Nutch

Nutch[25] es un software de búsqueda-web de código abierto. Está basado en LUCENE y en SOLR (el robot de búsqueda ha sido escrito desde cero exclusivamente para este proyecto), pero agregando características web específicas, tales como un crawler, una base de datos de enlace-gráfico, analizadores de HTML y otros formatos de documentos, etc. Nutch puede

[24] http://regain.sourceforge.net/

[25] http://nutch.apache.org/

ejecutarse en una máquina simple, pero se puede explotar mucho más su potencial si se ejecuta en un cluster Hadoop[26] (Nutch, 2010) (NutchWiki, 2010).

Características: (NutchWiki, 2010)

- Obtención, análisis e indexación en paralelo y/o distribuido.
- Plugins.
- Muchos Formatos: texto plano, HTML, XML, ZIP, OpenDocument, Microsoft Office (Word, Excel, Powerpoint), PDF, Javascript, RSS, RTF, MP3 (etiquetas ID3)
- Ontología (Recupera extensiones ontológicas)
- Clustering
- MapReduce [27]
- Sistema de fichero distribuido (a través de Hadoop)
- Base de datos de enlace-gráfico.
- Autenticación NTLM

1.5 Comparación y Selección de las Herramientas de Búsqueda

Una vez identificadas algunas herramientas de búsqueda de la sección anterior, en este punto se realiza una comparación entre ellas (véase tabla 1.1), de acurdo a los resultados obtenidos se procedió a seleccionar las herramientas que mejor se adapten a los intereses para un buscador de material educativo.

Para la comparación se tomaron en cuenta algunos criterios como: i) la herramienta de búsqueda a utilizar tiene que ser preferentemente de código abierto; para que la misma pueda ser utilizada sin ningún tipo de restricción ya sea para su uso directo o para alguna modificación; ii) la herramienta de búsqueda deberá poder utilizarse sin ningún inconveniente en distintas plataformas, iii) que sea multiplataforma, iv) el tipo de búsqueda que cada herramienta realiza y v) tipos de datos sobre los cuales trabaja.

En base al breve análisis de la [Tabla 1-1], se seleccionarán dos de las herramientas estudiadas; Watchdog queda descartada ya que es una aplicación que solo funciona con ontologías, así mismo Knowledge Hub y DiscoverEd son buscadores de OER en línea, solo siendo DiscoverEd de código abierto pero aun se encuentra en etapa beta. Del resto de herramientas se seleccionó Nutch y Regain para realizar el presente estudio. La herramienta Regain está basada en Lucene, y la herramienta Nutch utiliza Lucene y Solr, por lo que Lucene y SOLR quedan descartados.

Nutch es una herramienta que nos ofrece varias opciones, Regain ofrece solo búsquedas en texto plano, pero será tomado en cuenta en este estudio de tesis para realizar una

[26] El proyecto Apache Hadoop desarrolla software de código abierto para la informática fiable, escalable y distribuida. http://hadoop.apache.org/

[27] Permite a conjuntos de datos masivos ser procesados de forma distribuida, rompiendo la transformación en muchos cálculos pequeños.http://wiki.apache.org/nutch/MapReduce

comparación entre ambos buscadores y así determinar en qué grado los datos estructurados mejoran los resultados de búsqueda y también para determinar si una herramienta que realice búsquedas más complejas implica también un tratamiento especial en su configuración. Para una revisión más detallada acerca de estas dos herramientas véase el **Anexo 1**.

Tabla 1-1 Análisis de las herramientas candidatas.

	Watchdog	**Knowledge Hub**	**DiscoverEd**	**LUCENE**	**SOLR**	**SIREn**	**Nutch**	**Regain**
Tipo	Aplicación de escritorio	Aplicación web	Aplicación web	API	Plataforma búsqueda	plugin de Lucene	SRI - Basado en Lucene y Solr	Aplicación - Basado en Lucene
Multiplataforma	Si	Si	Si	Si	Si	Si	Si	Si
Código abierto	Si	NO	Si	Si	Si	Si	Si	Si
Lenguaje desarrollado	Java	-	Java	Java	Java	Java	Java	Java
Soporta documentos/ formatos	Ontologías	-	-	txt, pdf, doc, ppt, rtf, xml y html	-	Microformatos y RDF	tex, HTML, XML, ZIP, OpenDocument (OpenOffice. Org), Microsoft Office (Word, Excel, PowerPoint), PDF, Javascript, RSS, RTF, MP3 (ID3 tags)	txt, pdf, doc, ppt, rtf, xml y html
Búsqueda	-	-	-	indexado y búsqueda de texto completo	búsquedas de texto completo, resaltado de palabras importantes, búsqueda en facetas, agrupación dinámica, integración con bases de datos, y la manipulación de documentos (por ejemplo, Word, PDF)	Indexando y consulta RDF. Trabaja con documentos semi-estructurados (o documentos con esquemas muy diferentes).		Basado en Lucene
Observaciones					Contenedor de servlets Tomcat, Jetty, or Resin	SIREN utiliza la biblioteca de Lucene y SOLR, lo que significa que usando SIREn se pueden aprovechar las características de ambos	Integra apache SOLR, tomcat. Es motor de búsqueda libre. Tiene un carwler, una DB y un indexador	Búsquedas en texto plano

2. Evaluación del Sistema de Búsqueda

Propósito

Realizar una evaluación de las herramientas de búsqueda seleccionadas, tanto a nivel del usuario final como a nivel de rendimiento.

Introducción

Una vez estudiada la búsqueda para Recursos Educativos Abiertos y las herramientas de búsquedas para material educativo, en el presente capítulo se presenta la forma de evaluar dichos sistemas de búsqueda. La evaluación se considera desde 2 puntos de vista: i) evaluación a nivel del usuario final y ii) evaluación del rendimiento de las herramientas.

En la sección 2.1 se tratan los principales conceptos que intervienen en la evaluación del sistema de búsqueda, en la sección 2.2 se explica cómo se puede realizar la evaluación desde el punto de vista del usuario final; para finalmente, en la sección 2.3, estudiar la evaluación del rendimiento de los sistemas de búsqueda y la metodología que se siguió para lograr dicho propósito.

2.1 Definiciones

Antes de describir la metodología utilizada para comparar los sistemas de búsqueda, se presentan los conceptos más utilizados en la evaluación de los mismos:

Relevancia

A grandes rasgos, la relevancia mide la proximidad entre un documento y la formulación de una petición o expresión de la necesidad informativa. Basándose en el cálculo de relevancia, es posible determinar el valor de los índices de precisión y exhaustividad, indicadores de rendimiento de la recuperación (Amat, 2005). Para entender mejor el concepto de relevancia, (Martínez F. J., 2002) cita a (Cooper, 1973) quien introduce la idea de "utilidad de un documento", considerando que es mejor definir a la relevancia en términos de la percepción que un usuario posee ante un documento recuperado, es decir: "si el mismo le va a ser útil o no" (Cooper, 1973)**.** Con esta nueva perspectiva se puede asumir que un usuario tendrá problemas a la hora de definir qué es relevante y qué no lo es, pero tendrá pocos problemas a la hora de decidir si el documento le parece o no útil (Martínez F. J., 2002).

La precisión

La precisión mide el porcentaje de documentos recuperados que resultan relevantes con el tema de la pregunta y su cálculo es verdaderamente simple: se divide el total de documentos relevantes recuperados entre el total de documentos recuperados (Martínez F. J., 2002). Por ejemplo, suponiendo que un SRI contiene 40 documentos relevantes que satisfacen una consulta dada, y el sistema de recuperación solamente obtiene 30 documentos, de los cuales sólo 20 son relevantes; entonces la precisión del sistema es de 20/30, es decir 67% (Serna et al., 2004).

La exhaustividad

La exhaustividad mide la proporción de documentos relevantes que son recuperados. Corresponde al cociente entre el número de documentos relevantes recuperados y el total de documentos relevantes existentes en la colección (Amat, 2005). Por ejemplo, suponiendo que en la base de datos existen 40 documentos relevantes para la consulta de un usuario y que el sistema de recuperación obtiene 20 documentos relevantes, por lo tanto la exhaustividad es de 20/40, es decir 50% (Serna et al., 2004). El cálculo de la exhaustividad presenta algunas dificultades en los sistemas tradicionales, y más aún en los sistemas de recuperación en internet. En ambos casos, la determinación del denominador de la expresión (número total de documentos relevantes existentes en la colección) se estima comúnmente de forma indirecta. Se habla entonces de exhaustividad relativa (Amat, 2005).

Otros criterios de evaluación que se consideran son aquellos relacionados con la estructura de datos y algoritmos de recuperación; éstos son: la eficacia en la ejecución y la eficiencia del almacenamiento (Serna et al., 2004); estos criterios no serán evaluados en este estudio de tesis, puesto que los sistemas actuales ofrecen una respuesta muy rápida a las peticiones de los usuarios y más aún, la cantidad de documentos a recuperar es pequeña al tratarse de un

sistema que no realiza la búsqueda en toda la web; o en una gran parte de ella, sino en un determinado número de repositorios; a continuación se dará una breve explicación de estos criterios:

La eficacia en la ejecución es medida por el tiempo que toma un SRI para realizar una operación. Este parámetro es importante en un SRI, debido a que un largo tiempo de recuperación, interfiere con la utilidad del sistema, llegando a alejar a los usuarios del mismo si es lento (Serna et al., 2004).

La eficiencia del almacenamiento es medida por el número de bytes que se precisan para almacenar los datos. El espacio general, una medida común para medir la eficacia del almacenamiento, es la razón del tamaño del índice de los archivos más el tamaño de los archivos del documento sobre el tamaño de los archivos del documento (Serna et al., 2004).

2.1.1 Relación entre Precisión y Exhaustividad

La precisión y la exhaustividad tienden a relacionarse de forma inversa, ya que cuanto mayor es el valor de la precisión, menor va a mostrarse el valor de la exhaustividad (Martínez F. J., 2002).

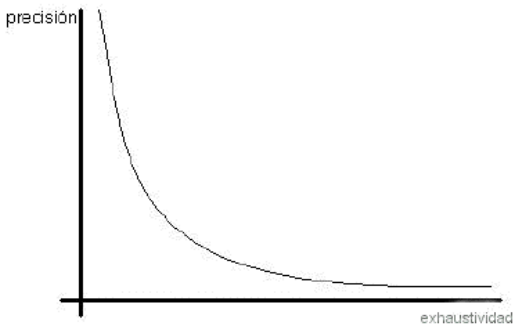

Figura 2.1 Evolución típica de la precisión y de la exhaustividad en un SRI (Martínez F. J., 2002).

Para entender mejor el enunciado expresado anteriormente, véase el **Anexo 8**, donde se considerará un ejemplo explicado en (Martínez F. J., 2002).

2.1.2 Cálculo de la Precisión y de la Exhaustividad

Si bien la precisión de una operación de recuperación de información puede ser calculada fácilmente, el cálculo de la exhaustividad se presenta inviable, "solamente puede ser estimado" (Blair, 1990). Ante esta situación se presentan dos enfoques (Martínez F. J., 2002):

1) Lo ideal sería analizar todos los documentos que se recuperen en una consulta e ir analizando uno por uno; pero por ejemplo, si tenemos un SRI cuya base de datos posee un tamaño de 1000 documentos, y un usuario recupera el 10% de los documentos de la base de datos (es decir, 100), a este usuario le puede resultar tedioso y aburrido analizar esta

cantidad de documentos uno a uno (para verificar la relevancia o no del mismo con la materia objeto de su necesidad de información), pero esa operación resulta posible.

Para un proceso mucho más preciso de esto, se debe conocer de antemano que documentos son relevantes para la consulta planteada; es decir, se conocerá el número exacto de documentos relevantes que el SRI posee para una pregunta dada; esto por supuesto resulta en un proceso manual y muy complicado en el que se deberán analizar todos los documentos existentes.

2) Otra opción y que parece mucho más lógica a la hora de evaluar un SRI muy extenso, es la idea de calcular la exhaustividad a partir de una muestra aleatoria de la colección documental, donde el usuario evaluará la pertinencia de los mismos y entonces, empleando técnicas estadísticas fiables, estimará el número de documentos útiles de la colección.

Los cálculos de la opción 1) resultan simples en el caso de que se de esta situación demasiado tediosa, y aunque sería lo ideal como se mencionó, no es un método muy utilizado; pero para la opción 2) es un proceso más complejo; aceptando que el cálculo de la precisión y exhaustividad debe llevarse a cabo sobre una muestra pequeña de la amplia colección de documentos de la base de datos, en (Martínez F. J., 2002) se expone cómo se realiza este cálculo, a continuación se explica un ejemplo de cómo realizar estos cálculos. En primer lugar, se supone que se elige una muestra constituida por los primeros ocho documentos (d1, d2,..., d8) recuperados en una búsqueda Q, en la que resultan pertinentes los documentos {d1, d2, d4, d6, d8}. Siguiendo lo indicado por (Salton & McGill, 1986), Francisco Martínez resume el proceso explicándolo de la siguiente manera: los valores de exhaustividad y precisión calculados son los siguientes:

Exhaustividad – Precisión			
N	Relevante	E	P
d1	X	0.2	1
d2	X	0.4	1
d3		0.4	0.66
d4	X	0.6	0.75
d5		0.6	0.60
d6	X	0.8	0.66
d7		0.8	0.57
d8	X	1	0.625

Figura 2.2 Cálculo de pares de valores E-P de la búsqueda de ejemplo (Martínez F. J., 2002).

Salton entiende que los cálculos Exhaustividad-Precisión (E-P), deben realizarse documento a documento recuperado, es decir, no son iguales el par de valores E-P en el primer documento que en el segundo. Cuando realizamos los cálculos en el primer documento (d1), se ha recuperado un único documento que es pertinente y, por tanto, la precisión va a valer uno (un acierto en un intento) y la exhaustividad (resultado de dividir el valor de uno entre el total de

documentos relevantes de la muestra, valor que sí conocemos de antemano y es cinco), vale 0.2. Así, el documento d1 tiene asignado el par de valores E-P (0.2, 1).

A continuación, se procede a calcular el par de valores E-P de d2, también relevante, aquí la precisión será el resultado de dividir el valor de dos documentos relevantes recuperados (d1 y d2) entre el total de documentos recuperados hasta el momento (dos también), por lo que adquiere de nuevo el valor de la unidad; la exhaustividad será el resultado de dividir el valor de dos (ambos son relevantes) entre el total de documentos relevantes de la muestra (cinco), obteniéndose un valor de 0.4, por lo que al documento d2 se le asignaría el par de valores E-P (0.4,1). Siguiendo este método se determinan el resto de los pares de valores E-P para los seis restantes documentos recuperados. Este conjunto de ocho pares de valores caracterizará, en principio, a la búsqueda Q (Martínez F. J., 2002).

2.2 Evaluación a Nivel de Usuario

En la siguiente sección se analizaron los requerimientos generales que un sistema de búsqueda debe tener; esto desde el punto de vista del usuario final, el mismo que usualmente no se preocupa del funcionamiento interno del sistema de búsqueda, sino que se fija en aspectos tales como: la interfaz gráfica, la facilidad de uso, la precisión de los resultados de búsqueda, etc.

2.2.1 Interfaz de Usuario

Como aspecto fundamental de un SRI está la interfaz de usuario, la interfaz juega un papel fundamental en el SRI ya que se trata del elemento catalizador entre el usuario y el sistema; hay que tener en cuenta que el principal objetivo perseguido por el SRI durante la búsqueda de la información es el de ofrecer resultados exhaustivos y precisos (Pastor & Artiga, 2009).

La interfaz de usuario ofrecerá facilidades al usuario a la hora de formular su consulta, ya que el usuario no tiene por qué saber exactamente el funcionamiento tanto externo como interno del sistema. También se ocupará de mostrar al usuario el resultado de búsqueda, una vez procesada su consulta (López, 2006).

Hay que tener en cuenta que la interfaz de usuario debe poseer una serie de características fundamentales. En (Asensi-Artiga, 1998) se cita a Tague y Schultz (1989) y a García Marco (1995) de la siguiente forma: “A este respecto, Tague y Schultz (1989) desarrollan los conceptos de Informatividad (capacidad de que el usuario pueda recibir información útil y no repetitiva de los registros consultados), amigabilidad (facilidad de acceso a las diversas funciones del SRI) y visualización de la recuperación (donde todos los elementos visualizados sean oportunos, adecuados y se presenten de un modo agradable). García Marco (1995) se centra en la temática de los interfaces amigables, detallando el significado de dicho término. Esencialmente indica que la interfaz amigable debe combinar distintos tipos de códigos comunicativos (visuales, auditivos, textuales), estructurar las funciones en niveles jerárquicos de fácil acceso al usuario, propiciar la inferencia metafórica por parte del usuario, ofreciéndole diversos tipos de ayuda durante la utilización de la interfaz.”

2.2.2 Sistema de Interrogación e Interfaz de Búsqueda

El sistema de interrogación toma las preguntas del usuario (generalmente estas preguntas se encuentran en lenguaje natural), elimina asiduosamente[28] las palabras no significativas comparando las ocurrencias de la consulta contra la lista de palabras vacías, y recorre el archivo invertido de la base de datos para seleccionar las entradas relevantes. La interfaz típica de un motor de búsqueda presenta una casilla para el ingreso de la ecuación de búsqueda y un botón para ejecutarla, otras veces además incorpora una casilla para optar como serán procesados los términos de la consulta. Pero la mayor parte de las opciones del sistema serán visibles cuando se haga uso de la búsqueda avanzada (Santesteban, 2001).

Algunas opciones de búsqueda avanzada incluyen (Santesteban, 2001):

- Búsqueda booleana y paréntesis
- Especificación de los términos que deben estar o no presentes
- Truncamiento (manual o filtrado por fecha, dominio, idioma, tipo de caracteres (normas ISO) o tipo de archivo (extensión)
- Búsqueda por retroalimentación
- Búsqueda sensible a letras mayúsculas

La mayoría de los motores de búsqueda, como respuesta a una consulta, presentan 10 (diez) o más resultados al mismo tiempo con un formato de visualización por defecto mostrando el título y algo de texto (Santesteban, 2001).

2.2.3 Recuperación y Consulta

El proceso que sigue a la búsqueda de información es la consulta de los elementos recuperados. Existen múltiples modos de visualización de los conjuntos recuperados en general y de cada elemento en particular. Quizá sea éste el momento más crítico durante el uso de los SRI ya que de esto dependerá en gran parte la aceptación del usuario hacia el sistema. Cada conjunto de elementos recuperados debe visualizarse en una ventana con el contenido de algunos campos esenciales como son título, año, autor, fuente, etc. Para evitar la confusión del usuario, debido a la proliferación de ventanas, la interfaz debe presentar una ventana con un histórico de las consultas realizadas y el número de elementos recuperados; las distintas consultas deberán llevar algún tipo de marca o icono identificativo, que informe al usuario sobre si una determinada consulta se trata de un agente en progreso y si se ha guardado como estrategia de búsqueda. Una tercera ventana mostrará el contenido del elemento seleccionado en la ventana de elementos recuperados. La ventana de resultados debe exponer de forma clara y estructurada los elementos recuperados. De este modo se evita que el usuario pueda sufrir una desorientación o un desbordamiento cognoscitivo[29], debido fundamentalmente a la gran cantidad de ítems que pueden recuperarse durante el proceso de búsqueda (Asensi-Artiga, 1998).

[28] Actividad que se hace constante y frecuentemente

[29] Gran cantidad de operaciones y decisiones que el usuario enfrenta a la vez.

2.2.4 Evaluación de los Sistemas de Recuperación de Información

En este punto se determinan algunos requerimientos del sistema, partiendo desde cómo se evalúan los SRI, es decir mediante una retroalimentación.

Un SRI puede evaluarse empleando diversos criterios. (López, 2006) cita a Frakes, quien selecciona los dos siguientes como los más importantes: i) ejecución eficaz (eficacia). La importancia relativa de estos factores debe decidirla el diseñador del sistema, y ii) la selección de la estructura de datos y los algoritmos apropiados para su implementación dependería de esa decisión. La eficacia en la ejecución se medirá por el tiempo que toma el sistema o una parte del mismo para llevar a cabo una operación. Este parámetro ha sido siempre una preocupación principal en un SRI, especialmente desde que muchos de ellos son interactivos y un tiempo de recuperación excesivo interfiere con la utilidad del sistema, llegando a alejar a los usuarios del sistema. Los requerimientos no funcionales de un SRI normalmente especifican el tiempo máximo aceptable para una búsqueda y para las operaciones de mantenimiento de una base documental, tales como añadir y borrar documentos. La eficiencia del almacenamiento se medirá por el número de bytes que se precisan para almacenar los datos. El espacio general, una forma común de medir la eficacia del almacenamiento, es la razón del tamaño de los ficheros índice más el tamaño de los archivos del documento sobre el tamaño de los archivos del documento. Tradicionalmente, se le ha dado mucha importancia a la efectividad de la recuperación, normalmente basada en la relevancia de los documentos recuperados a las necesidades reales de información del usuario, lo cual ha representado un problema ya que medir la relevancia es un proceso subjetivo y sin confianza. Esto es, diferentes juicios personales asignarían diferentes valores de relevancia a un documento recuperado en respuesta a una búsqueda.

Por otro lado, en (López, 2006) también se hace referencia a Salton y McGill, quienes señalan que, además de los criterios anteriores que se centran principalmente en el punto de vista del diseñador del sistema, se debe considerar también el punto de vista del usuario ya que los criterios de evaluación del diseñador y del usuario no tienen por qué coincidir. Los seis criterios siguientes han sido identificados como los más importantes en lo que respecta a las características que un SRI debe ofrecer al usuario:

1. La exhaustividad, o habilidad del sistema para presentar todos los ítems relevantes.
2. La precisión, o habilidad del sistema para presentar solamente ítems relevantes.
3. El esfuerzo, intelectual o físico, requerido por el usuario en la formulación de las consultas, en el manejo de la búsqueda y en el proceso de examinar los resultados.
4. El intervalo de tiempo transcurrido entre que el sistema recibe la consulta del usuario y presenta las respuestas.
5. La forma de presentación de los resultados de la búsqueda, la cual influye en la habilidad del usuario para utilizar la información recuperada.
6. El alcance o cobertura de la colección documental, o la proporción en la que están incluidos en la recuperación todos los ítems relevantes del sistema ya conocidos por el usuario.

Otros requerimientos que no han sido nombrados hasta el momento son algunos requerimientos no funcionales[30]:

- Integridad
  La información sólo puede ser modificada por quien está autorizado y de manera controlada.
- Disponibilidad
  El SRI debe estar disponible cuando se le necesite.
- Confiablidad
  Se refiere a proporcionar la información apropiada

2.2.5 Características Generales del Sistema de Búsqueda para OERs

Aunque para el usuario final el proceso de búsqueda es transparente, en la siguiente sección se hace mención a las tecnologías semánticas con el fin de mostrar al lector la diferencia de búsqueda entre ambas herramientas de seleccionadas.

El sistema de búsqueda para OERs, además de contar con los requerimientos nombrados en la sección anterior, deberá tener soporte para tecnologías semánticas como son por ejemplo: RDF, RDFa, microformatos, RSS o Atom; las mismas que fueron estudiadas en la sección 1.2.7

2.2.6 Resumen de Criterios y Aspectos para la Evaluación a Nivel de Usuario

Una vez identificados los aspectos que se deben considerar para elegir un sistema de búsqueda, se describen las métricas que se pueden utilizar para evaluarlo. A continuación, se expone una matriz en la que se resumen los diferentes requerimientos y su respectiva calificación para las dos herramientas de búsqueda seleccionadas para este estudio de tesis.

[30] Tienen que ver con características que de una u otra forma puedan limitar el sistema

Tabla 2-1 Evaluación del Sistema de búsqueda para OER.

	Criterio	Descripción
Requerimientos	Amigabilidad	facilidad de acceso a las diversas funciones del SRI
	Informatividad	capacidad de que el usuario pueda recibir información útil y no repetitiva de los registros consultados
	Visualización de recuperación	todos los elementos visualizados sean oportunos, adecuados y se presenten de un modo agradable
	Búsqueda avanzada	Búsqueda a mas detalle
	Eficacia	el tiempo que toma el sistema o una parte del mismo para llevar a cabo una operación
	Eficiencia	Consiste en que la información sea generada con el óptimo (más productivo y económico) uso de los recursos.
	Integridad	La información sólo puede ser modificada por quien está autorizado y de manera controlada.
	Disponibilidad	El SRI debe estar disponible cuando se le necesite.
	Confiabilidad	se refiere a proporcionar la información apropiada
	Exhaustividad	habilidad del sistema para presentar todos los ítems relevantes
	La precisión	o habilidad del sistema para presentar solamente ítems relevantes
Tecnologías semánticas	Metadatos	Trabaja con los metadatos
	Rss, Atom, Feeds	
	RDF	Soporte para RDF
	RDFa	Soporte para RDFa
	Microformatos	Soporte para microformatos

En el capítulo 5 se realiza una evaluacion de las 2 herramientas seleccionadas en base a los criterios de la [Tabla 2.1].

2.3 Evaluación del Rendimiento

Cuando se accede a un buscador, el usuario normalmente encuentra una página web que presenta una plantilla o formulario en la que introduce la ecuación de búsqueda constituida por palabras clave, operadores booleanos, etc., donde se han de encontrar los términos en el documento y demás datos que se consideren necesarios para delimitar y centrar la consulta. Una vez procesada, el buscador muestra los resultados ordenados según su relevancia probable relativa a la pregunta planteada. Estos resultados serán el principio para realizar la evaluación; el método utilizado para la evaluación del Sistema de Búsqueda es el propuesto por (Olvera, 2000) para la evaluación de herramientas de recuperación de información; el cual se puede aplicar tanto en buscadores generales así como en buscadores especializados,

aunque se realizarán algunas variantes a la hora de calcular la relevancia de los documentos; el método de evaluación consta de 5 etapas principales (Olvera, 2000):

a) Determinación de las necesidades de información de los usuarios.
b) Elaboración del enunciado de búsqueda.
c) Realización de las consultas.
d) Valoración de la relevancia.
e) Análisis de los resultados.

El proceso de evaluación se inicia con la elaboración de las ecuaciones de búsqueda mediante la sintaxis correspondiente a partir de las necesidades de información planteadas el usuario; tras realizar las consultas en el buscador, se valora la relevancia de los ítems recuperados y finalmente, se analizan los resultados conforme a las medidas de exhaustividad y precisión. A continuación se describen los pasos a realizar para la aplicación del método de evaluación del SRI:

a) Determinación de las necesidades de información de los usuarios.

La determinación de las necesidades de información de los usuarios es la primera fase a llevarse a cabo para la evaluación de un SRI, normalmente son los propios investigadores los que proponían las preguntas para interrogar al sistema, pero esto puede conllevar a un sesgo y a una falta de imparcialidad; al menos, potencial. Aunque importantes proyectos en el tema como la Conferencia de Recuperación de Textos (TREC) usan la colaboración de asesores externos para realizar esta tarea; la tendencia en la evaluación del funcionamiento de los SRI en internet se orienta a recoger preguntas del servicio de referencia de las bibliotecas o de estudiantes; es decir, de usuarios reales de información (Olvera, 2000).

La selección de las preguntas es un aspecto clave, ya que, en gran medida, de ella depende el éxito o fracaso de la prueba. Las preguntas ofrecen el punto de partida para realizar las consultas, para controlar el proceso de búsqueda y para valorar los resultados ofrecidos por el sistema. Las preguntas deberían presentar las siguientes características (Olvera, 2000):

- Que sean preguntas que, muy probablemente, se encuentren en los repositorios seleccionados.
- Que constituyan una combinación de preguntas 'fáciles' con un alto nivel de respuesta y 'difíciles' con resultados más restringidos en relación a la cantidad de recursos que sobre ellas se pudieran encontrar
- Que unas preguntas sean de temas académicos y/o especializados y otras de temas más comunes y que se trate de preguntas heterogéneas, relacionadas con temas diversos.

En (Olvera, 2000) se menciona que muchos investigadores sobre este tema proponen en sus trabajos un número diferente de preguntas que deberían realizarse; es decir, que no existe un acuerdo común referente al número de preguntas a utilizar; Olvera utiliza un total de 20 preguntas en su investigación.

b) Elaboración del enunciado de búsqueda (La sintaxis de búsqueda).

El reto principal al realizar una consulta es conseguir que la pregunta recupere los documentos que se consideran realmente relevantes. Para realizar las consultas en los SRI, las preguntas son traducidas a las expresiones o enunciados de búsqueda correspondientes. Dicha expresión de búsqueda puede constar de varios elementos: términos, operadores lógicos, uso de paréntesis, truncamiento, formulación de la búsqueda en lenguaje natural, etc. La naturaleza cambiante de las preguntas demanda sintaxis de búsqueda diferentes, ya sean expresiones booleanas, de frase, de un término, etc.; y se ha de escoger la que en cada caso resulte, probable intuitivamente, más adecuada sin descuidar que se ha de contribuir a la homogeneidad de los resultados para facilitar su comparación (Olvera, 2000).

Como ejemplo de una expresión de búsqueda representada de diferentes formas tenemos (Felquer et al., 2001):

- Bibliotecas Escolares
- "bibliotecas escolares"
- +bibliotecas+ escolares

c) Realización de las consultas.

En esta fase se plantean las preguntas en todos los servicios de búsqueda a analizar.

d) Valoración de la relevancia.

En este paso se determina la relevancia de cada documento recuperado. La mayor parte de los buscadores ordenan los resultados en función de su relevancia respecto a la pregunta planteada. Por lo general, los mejores resultados aparecen siempre en la parte superior de la lista de referencias. Ya que al ser un SRI que solo realiza la búsqueda en determinados repositorios, la cantidad de elementos recuperados será mínima en comparación con un buscador tradicional cuyo índice contendrá cientos o miles de sitios web en sus registros. Por esta razón, las consultas arrojarán pocos resultados o incluso ninguno si las palabras clave adecuadas no se usan. Para el análisis de la relevancia se seguirá el método 1) descrito en el capítulo II sección 2.1.2 ‘Cálculo de la precisión y de la exhaustividad’; en el cual se analizan todos los elementos de los repositorios de forma manual, particularmente en este caso se analizará una categoría en común de todos los repositorios para ver cuáles son relevantes con respecto a la pregunta; la categoría seleccionada es la de ‘Anthropology’, esta categoría se seleccionó en base a la [Tabla 5-5] ‘Número de recursos por categoría y repositorio’. Para evaluar la relevancia se utiliza una escala constituida por varias categorías:

Tabla 2-2 Escala de evaluación para medir la relevancia (Olvera, 2000).

RELEVANCIA	0: Enlaces **duplicados, inactivos e irrelevantes** (que no satisface la pregunta ni recoge los términos de la ecuación de búsqueda)
	1: Enlaces técnicamente **adecuados pero no útiles** (que recogen las diferentes partes de la pregunta pero no en el contexto adecuado; el documento puede contener los términos o componentes de la pregunta, pero bastante alejados entre sí. También se asigna 1 punto a páginas que mencionan el tema en el contexto adecuado pero que sólo contienen un mínimo de información realmente útil)
	2: Enlaces potencialmente **útiles** (que no abordan el tema en profundidad o se centran en algún aspecto específico del mismo, o páginas con al menos un enlace a otra página a la que se le asignan 3 puntos)
	3: Enlaces probablemente **más útiles** (que tratan el tema extensamente y serán útiles para quien plantee la pregunta)

A continuación se describe con más detalle la escala de relevancia (Olvera, 2000):

Duplicados

Si el enlace en cuestión tiene el mismo URL (Uniform Resource Locator) básico que un enlace anterior de la lista de resultados, se lo considera en la categoría de duplicados, independientemente de sus otras cualidades (inactivo, irrelevante o válido). Esta categoría incluye variantes muy obvias pero otras son más sutiles: si un nombre del directorio en el URL está en mayúsculas en un caso pero no en otro, cuenta como duplicado. Los espejos (mirror sites o alias), servidores idénticos que tienen direcciones IP o directorios diferentes, incluso cuando dos archivos son el mismo o versiones ligeramente diferentes, no se consideran como duplicados. Si dos o más enlaces hacen referencia a un mismo elemento el cual pertenezca a más de una categoría, se toma al primer elemento como si fuese único y el resto serán tomados cómo duplicados.

Inactivos

Se consideran enlaces inactivos los que se encuentran entre los casos siguientes:

- Error 404: el servidor ha sido contactado pero no se consigue localizar ese fichero.
- Error 603: el servidor no responde, para los errores 404 y 603 se comprueban los enlaces varias veces, por ejemplo, en un periodo de una semana.
- Mensajes que indican que el acceso a la página está prohibido o que se necesita clave de acceso.
- Mensajes que anuncian que la página deseada ha sido eliminada o trasladada a otro servidor.

e) Análisis de los resultados.

Análisis de los resultados obtenidos.

El proceso para la evaluación del rendimiento de los sistemas de búsqueda se observa en la Figura 2.3:

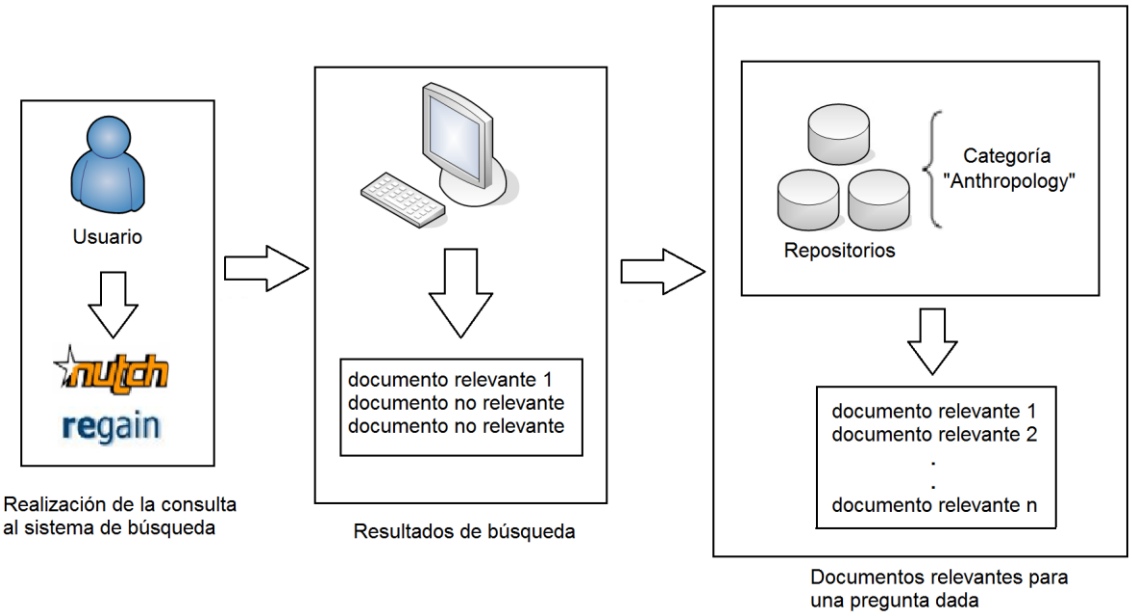

Figura 2.3 Proceso para evaluar el rendimiento del sistema de búsqueda.

Primero el usuario realiza la consulta al sistema de búsqueda de recursos educativos abiertos; a continuación el sistema de búsqueda arroja los resultados correspondientes a dicha pregunta, que podrán ser documentos relevantes o no relevantes; luego de esto, como se observa el la figura 2.3, solo se ha realizado la búsqueda sobre una determinada categoría que corresponde a la categoría de antropología y debido a que se realizó un análisis manual de todos los recursos pertenecientes a esta categoría en todos los repositorios, se sabe de antemano que documentos son relevantes frente a una pregunta determinada.

En el Anexo 8.5 se describe brevemente cada recurso educativo perteneciente a la categoría de antropología para cada uno de los repositorios seleccionados.

3.Análisis y Selección de Fuentes de Material Educativo para la Búsqueda

Propósito

Determinar los criterios de evaluación para la selección de los repositorios a utilizar e identificar los repositorios que mejor se adapten a estas características para la implementación del piloto de búsqueda.

Introducción

El internet es un vasto mar de información, pero también de información basura (Craik, 2010), es por esto que la elección de los repositorios a considerar para el Sistema de Recuperación de Información es de vital importancia.

Se necesita usar la misma evaluación crítica en la búsqueda de información en internet que se haría para un libro, un índice de papel, una partitura musical, o en una base de datos comercial en línea. El contenido de internet es solo más diverso debido a su potencial interacción con demás medios de comunicación. Con el crecimiento de información en internet y el desarrollo de herramientas de búsqueda, existe mayor posibilidad de encontrar la información que se busca en el internet, pero a la vez esta información valiosa es muy escasa en un mar de información basura (Tillman, 2003). Los bibliotecarios pueden evaluar las fuentes de información de Internet para juzgar la calidad o la pertinencia de la información para una consulta en particular o usuario (Smith, 1997).

El potencial de Internet como fuente de ingresos importante, sólo puede realizarse si los usuarios pueden encontrar exactamente lo que quieren de manera rápida, precisa y con poco esfuerzo. El desarrollo de herramientas que apoyan el hallazgo de material pertinente, dentro de pocos clics de ratón y pulsaciones de teclas se está convirtiendo cada vez más críticos, dado el ritmo sin precedentes en la que la web está creciendo y el número creciente de usuarios novatos (Pollock & Hockley, 1997).

Los repositorios pueden albergar todo tipo de materiales siempre que sea posible su expresión en forma digital, por eso, pueden ser objeto de depósito desde ficheros textuales a audiovisuales, esto; si la tecnología del mismo lo permite. El hecho de albergar los objetos de aprendizaje en un repositorio, no garantiza su calidad; en los repositorios existen distintos tipos de materiales como son los materiales docentes, conferencias invitadas, informes, etc., que no pasan por mecanismos de evaluación (creación, certificación y comunicación) y, por lo tanto, será el usuario lector quien juzgue la calidad de los mismos; esta calidad a veces medida por el uso que se hace de ellos(accesos, descargas, citas, etc.) (BiD, 2008).

En el presente capítulo se analizará cómo evaluar dichos repositorios a nivel de contenidos, es decir, qué atributos clave, indicadores de calidad, etc. se consideran necesarios para la elección de los repositorios y de acuerdo a esto seleccionar los que mejor se adapten a las características planteadas, de manera que el sistema de recuperación a implementar en este proyecto, trabaje sobre un corpus de material educativo de fuentes importantes y así los resultados que se obtengan y se presenten al público sean más relevantes.

3.1 Búsqueda y Evaluación de los Recursos Web

El crecimiento exponencial y heterogéneo de recursos electrónicos educativos en Internet, provenientes tanto de profesores, asociaciones de estudiantes, bibliotecarios, empresas e-learning, etc., hacen que el usuario se sienta naufrago y desconcertado sin saber qué opciones seleccionar, pues no dispone de información fiable que le indique la calidad de los recursos, atendiendo a aspectos tan importantes como el contenido, la función y la usabilidad. Hay que tener en cuenta que muchas personas tienen dificultad para encontrar la información que necesitan, y en muchos casos no están seguros sobre la calidad de la información que encuentran o la credibilidad de las fuentes que consultan (Pinto, 2008).

La evaluación en este ámbito será orientadora, promotora de la calidad, garantizadora de la accesibilidad de los materiales, facilitadora de una información actualizada y analítica sobre los materiales multimedia existentes en el mercado. No toda la información que se genera y a la que se tiene acceso a través de Internet es igual de buena, ni igual de útil, ni igual de válida. La capacidad de evaluar y discernir la buena de la mala información para su uso posterior determinará el éxito del individuo o del grupo en este nuevo entorno en el que la información se ha hecho tan valiosa (Pinto, 2008). “La calidad de la información de un recurso educativo electrónico vendrá determinada por su capacidad para satisfacer las necesidades de información/formación de los estudiantes y profesores que lo utilicen o consulten” (Pinto, 2008). Pero esto es relativo, puesto que lo que para una persona puede ser válido, puede no serlo para otra.

3.1.1 Indicadores Clave de la Calidad

En (Tillman, 2003) se mencionan los siguientes indicadores de calidad; cabe recalcar que para comprender estos indicadores generalmente se lo realiza en forma de preguntas.

- La facilidad de conocer el alcance y los criterios de inclusión que permitirán ver si hay algún recurso de acuerdo a las necesidades del usuario.
- Facilidad de identificar:
  - la autoridad de los autores
  - Circulación/concurrencia
  - la última actualización
  - lo que se actualizó
- La estabilidad de la información:
  - ¿Se puede confiar en la información que permanece en dicho repositorio?
- Facilidad de uso, tanto en términos de conveniencia y de velocidad de conexión
  - Por ejemplo, si alguien ha puesto una gran película gráfica o QuickTime, ¿valdrá la pena esperar mientras se descarga? Esto no es diferente a que si una publicación está en un idioma que no sea la lengua natal del usuario, ¿valdrá la pena el esfuerzo realizado de traducir dicho documento?

3.1.2 Criterios de Calidad y Evaluación de la información en Internet

En (Smith, 1997), (The Virtual Chase, 2008) y (Beck, 1997) se mencionan una serie de criterios para la evaluación, selección, revisión, o calificación de los sitios de internet. Algunos de los criterios nombrados, responden a una o varias preguntas planteadas:

- **Alcance de la cobertura**

Se refiere al grado en que una fuente explora un tema, considera periodos de tiempo, geografía y jurisdicción, y la cobertura de temas relacionados.

- ¿Qué temas se tratan?
- ¿Qué ofrece esta página que no ofrecen otros lugares?
- ¿Cuál es su valor intrínseco[31]?
- ¿Cuán profundo es el material?

Justificación

1. La cobertura web a menudo difiere de la cobertura de impresión.
2. Con frecuencia, es difícil determinar la extensión de la cobertura de un tema de una página web. La página puede o no puede incluir enlaces a otras páginas web o referencias de impresión.
3. A veces la información web es "sólo por diversión", un engaño.
4.

- **Autoridad**

La autoridad se refiere a la experiencia o al estatus de reconocimiento oficial de una fuente. Hay que tener en cuenta la reputación del autor y el editor. Cuando se trabaja con información del gobierno, se debe considerar si la fuente es el proveedor oficial de la información.

- ¿Hay un autor?
- ¿El autor es calificado? ¿Es un experto?
- ¿Quién es el patrocinador?
- ¿El patrocinador de la página es de buena reputación? ¿Qué tan buena?
- ¿Hay un enlace a la información sobre el autor o el patrocinador?
- Si la página no incluye ni una firma, ni indica un patrocinador, ¿Hay alguna otra manera de determinar si origen?

Justificación

1. Cualquier persona puede publicar cualquier cosa en la web.
2. A menudo es difícil determinar la autoría de una página web.
3. Incluso si una página ha sido firmada, las calificaciones no son proporcionadas generalmente.
4. El patrocinio no siempre se indica.

[31] Valor o utilidad inherente a alguna cosa, independientemente de si sirve para satisfacer necesidades y aspiraciones del ser humano.

- **Objetividad**

La objetividad es la tendencia a la opinión expresada cuando un escritor interpreta o analiza los hechos. Se debe considerar el uso de un lenguaje persuasivo, la presentación de la fuente desde otros puntos de vista.

- ¿La información muestra un mínimo de sesgo[32]?
- ¿Está la página diseñada para influir en la opinión?
- ¿Hay alguna publicidad en la página?

Justificación

1. Con frecuencia los objetivos de los patrocinadores / autores no están claramente establecidos.

- **Precisión**

Describe la información que resulta incontestable y completa. ¿Es la información precisa en el recurso? Un recurso puede comprobarse contra otros recursos o con la información que el evaluador tiene.

- ¿La información es fiable y libre de errores?
- ¿Existe un editor o alguien que verifique/compruebe la información?

Justificación

1. Cualquier persona puede publicar cualquier cosa en la web.
2. A diferencia de los recursos de impresión tradicionales, los recursos web rara vez tienen editores o verificadores de datos.
3. Actualmente, no existen estándares web para asegurar la exactitud.

- **Actualidad**

Se refiere a la información que está vigente a la fecha de publicación

- ¿Está la página actualizada?
- En caso afirmativo, ¿Cuándo fue la última actualización?
- ¿Cómo son los enlaces actuales? ¿Algunos se han quitado o movido?

Justificación

1. La fecha de publicación o revisión no siempre se ofrece.
2. Si se proporciona una fecha, puede tener varios significados. Por ejemplo:
   - Puede indicar cuando fue escrito el material
   - Puede indicar cuando fue colocado el material en la web
   - Puede indicar cuando fue revisado por última vez el material

[32] Es la diferencia entre el valor esperado de un estimador y el verdadero valor del parámetro. El no tener sesgo es una propiedad deseable de los estimadores

- **Amplitud**
  ¿Qué aspectos de la asignatura están cubiertos? ¿El recurso se centra en un área específica o incluye temas relacionados?
- **Profundidad**
  ¿Cuál es el grado de detalle sobre el tema? Esto se relaciona con el nivel de audiencia para el cual ha sido diseñado el recurso.
- **Tiempo**
  ¿Es la información del recurso limitada por ciertos periodos de tiempo?
- **Contenido**
  ¿Es la información un hecho o una opinión? ¿El sitio contiene información original o solamente enlaces? Los usuarios pueden verse frustrado por las listas de recursos que parecen prometedores, pero resultan simplemente más enlaces. ¿El recurso es independiente, o ha sido extraído de otra fuente, quizás perdiendo el significado o enlaces en el proceso?
- **Fuentes**
  ¿Quién es el autor? ¿La fuente es creíble? No crea en todo lo que lee.
- **Unicidad/singularidad**
  ¿El contenido del recurso está disponible en otras formas (en otros sitios, en forma impresa, CD-ROM, etc.)? ¿Qué ventajas tiene este recurso en concreto? ¿Si el recurso se deriva de otro formato, tiene todas las características del original? ¿Se han añadido características extra?
- **Enlaces**
  Si el valor del sitio radica en sus enlaces hacia otros recursos, ¿son esos vínculos correctos, se mantienen actualizados? ¿Hay posibles problemas con los derechos de autor?
- **Escritura**
  ¿Está el texto bien escrito? Mientras la vinculación de multimedia e hipertexto son elementos importantes de la Web, la mayor parte del contenido de la información en la web todavía se encuentra en el texto, y la calidad de la escritura es importante para que el contenido sea claramente comunicado.
- **Diseño gráfico y multimedia**
  ¿Es el recurso interesante a la vista? ¿Los efectos visuales mejoran el recurso, distraen del contenido o son un sustituto del contenido? ¿El audio, vídeo, modelado de realidad virtual, u otros efectos que se utilizan, son adecuados para el propósito de la fuente?
- **Propósito y audiencia**
  ¿Cuál es la finalidad del recurso? ¿Está claramente establecido? ¿El recurso cumple con los objetivos planteados?
  ¿Quiénes son los destinatarios de este recurso? ¿El recurso satisface las necesidades de los usuarios previstos?
- **Comentarios**
  ¿Qué dicen los comentarios acerca del sitio? Esto ayuda a familiarizarse con las fortalezas y debilidades de las herramientas de revisión de recursos de Internet.
- **Trabajabilidad**
  ¿Es el recurso conveniente, y puede ser utilizado de manera efectiva?

La trabajabilidad es el área donde los criterios para los recursos de internet se diferencian más de las fuentes de impresión.

Un problema en el acceso a los documentos electrónicos es si una biblioteca debe proporcionar enlaces al sitio de origen o adquirir la publicación para el acceso local. Una trabajabilidad pobre puede indicar que la biblioteca debe almacenar los datos localmente, si la propiedad intelectual permite esto.

- **Ergonomía**
  ¿Es el recurso fácil de usar? ¿Es necesario algún comando especial? ¿Información de ayuda está disponible? ¿En la interfaz de usuario se han abordado cuestiones tales como el diseño del menú y la legibilidad de las pantallas?
- **Necesidades informáticas**
  ¿Se puede acceder al recurso con equipos y software estándar, o hay un software especial, contraseña o requisitos de la red? Es útil para poner a prueba los recursos con una variedad de navegadores y conexiones
- **Búsqueda**
  ¿Con qué eficacia puede ser recuperada la información de los recursos?
- **Navegabilidad**
  ¿Están los recursos organizados de una manera lógica para facilitar la localización de la información?
- **Interactividad**
  ¿Las características interactivas tales como formularios o scripts funcionan correctamente, tienen un valor añadido al sitio?
- **Conectividad**
  ¿Se puede acceder a los recursos de manera fiable?, ¿es la conexión de ancho de banda limitada?
- **Costo**
  En la actualidad, la información sobre los recursos de Internet es percibida como libre. Sin embargo, los costos sí existen, y es probable que se vuelvan más importantes. Los costos pueden dividirse en: (1) costes de conexión a los recursos, y (2) los costos asociados con el uso de la propiedad intelectual contenida en el recurso.

En un estudio realizado por (Smith, 1997) se analizaron 10 sitios web sobre los cuales se hizo la evaluación, mediante un checklist se analizaron los criterios para la evaluación de recursos de información en internet; y aunque el estudio se realizó sobre 26 de los criterios descritos anteriormente, al final de esta investigación se señala que solo 9 atributos tienen mayor relevancia, y una de las conclusiones que se dan es que: “La apariencia es ampliamente considerada como importante, incluso entre los sitios que están principalmente relacionados con el contenido. La organización del sitio y la facilidad con que los usuarios pueden encontrar su camino también se consideran importantes en el entorno de Internet. Todos los sitios de evaluación incluyen algún aspecto del contenido y la aplicabilidad. Los criterios de referencias tradicionales de los bibliotecarios como la moneda, autoridad, y la audiencia son también ampliamente utilizados.” (Smith, 1997) Estos criterios se pueden observar en la figura siguiente:

Tabla 3-1 Frecuencia de aparición de los criterios (Smith, 1997).

CRITERIOS	NÚMERO DE SITIOS
Diseño gráfico y multimedia	10
Navegabilidad y organización	8
Moneda	8
Contenido (en general)	7
Autoridad	5
Unicidad	4
Audiencia	4
Trabajabilidad (en general)	4
Conectividad	4

Esto da una indicación del tipo de criterios que los selectores, revisores y servicios de calificación sienten como importantes a la hora de trabajar o/no con algún sitio en específico. Los criterios de la [Tabla 3-1] se muestran en orden de la frecuencia con que los sitios de evaluación de las mismas se mencionan.

3.2 Criterios a Considerarse para la Selección de los Repositorios

En esta sección se analiza seis repositorios OER para realizar la búsqueda en ellos; se trata de repositorios heterogéneos, cuatro de los cuales son de propósito general como son: MIT, Connexions, OER Commons y Merlot; de igual forma dos repositorios más que ofrecen contenido audiovisual como son Uchannel y Edutube. Repositorios heterogéneos en cuanto al contenido de cada repositorio; es decir, qué tipo de recursos se albergan en los repositorios.

Se ha realizado un listado de los requerimientos que deberían cumplir los repositorios desde los cuales se realizará la extracción de la información para alimentar el Sistema de Recuperación; los requerimientos aquí expuestos, sin embargo, no toman en cuenta requerimientos como el diseño gráfico y multimedia, la navegabilidad y organización como se puede apreciar en la [Tabla 3-1], ya que para realizar la extracción de los recursos no se tomará en cuenta la amigabilidad del sitio, sino la disposición y calidad de los contenidos del mismo. Como resultado de ello se ha elaborado un checklist con los criterios que deberían cumplir los sitios a ser utilizados para la extracción de los recursos.

Estos criterios han sido seleccionados en base a las necesidades de este proyecto y son una abstracción de las secciones 3.1.1 y 3.1.2.

Tabla 3-2 Criterios para la selección de los repositorios.

		OPENCOURSEWARE CONSORTIUM (MIT)	CONNEXIONS	OER Commons	Uchannel	EduTube	Merlot
Criterios calidad del sitio/información							
Autoridad							
	¿Hay un autor?	1	1	1	0	0	1
	¿El autor es calificado?	1	1	1	0	0	1
	¿Hay como contactarlo?	0	1	0.5	0	0	1
	¿Patrocinador de la página es de buena reputación?	1	1	1	1	1	1
Alcance							
	¿Se trata el tema específico?	1	1	1	1	1	1
	Valor intrínseco	1	1	0.5	1	1	1
	Profundidad del material	1	1	0.5	1	1	1
Objetividad							
	No hay publicidad en la página	1	1	1	1	1	1
	¿No está la página diseñada para influir en la opinión?	1	1	1	1	1	1
Precisión							
	¿La información es fiable?	1	1	1	1	0.5	1
	¿Existe un editor o alguien que verifique/compruebe la información?	1	0	-	-	0	-
Actualidad							
	¿Está la página actualizada?	1	1	1	1	1	1
	¿Los enlaces funcionan correctamente?	1	1	1	1	1	1
Contenido							
	¿Recursos es independiente?	1	1	0.5	1	1	1
	¿La información es un hecho?	1	1	1	0.5	0.5	1

Tabla 3-2 (Continuación) Criterios para la selección de los repositorios

		OPENCOURSEWARE CONSORTIUM (MIT)	CONNEXIONS	OER Commons	Uchannel	EduTube	Merlot
Criterios calidad del sitio/información							
	¿El sitio contiene información original? (1) o solamente enlaces? (0)	1	1	0.5	1	1	1
Amplitud							
	¿El curso se centra en un área específica?	1	1	1	1	1	1
Profundidad							
	¿Se especifica el grado de audiencia?	1	0.5	1	0	0.5	1
Trabajabilidad							
	¿Es el recurso conveniente?	1	1	0.5	0.5	1	1
	¿Puede ser utilizado de manera efectiva?	1	0.5	0.5	0.5	1	1
TOTAL		20/20	18/20	15/20	13.5/20	14.5/20	20/20
Características de extracción							
RDF		x		x			
RDFa							
RSS		X (1.0)		? 1.0	x 2.0	x 2.0	x 2.0
Xml		x	x	x	x	x	x
Atom			X				
Otros							
¿Hay cómo extraer la información?							
Si		x	x	x	x	x	x
No							

En la siguiente tabla se presenta un resumen de la calificación obtenida por cada repositorio:

Tabla 3-3 Calificación obtenida para los repositorios estudiados

Repositorios	Tipo	Calificación /20
Uchannel	Audiovisual	13,5
EduTube	Audiovisual	14,5
OER Commons	Propósito general	15
Connexions	Propósito general	18
Merlot	Propósito general	20
Open Courseware Consortium (MIT)	Propósito general	20

Como se puede observar en la [Tabla 3-3], los repositorios OER Commons y Uchannel obtuvieron un puntaje inferior dentro de su categoría; es decir, repositorios de propósito general y audiovisual respectivamente; por lo que sobre estos dos repositorios no se realizará la búsqueda de información. En el repositorio Uchannel no se proporciona información acerca del autor del recurso, requisito fundamental al tratarse de un recurso educativo abierto; y aunque el repositorio OER Commons en el análisis individual de cada parámetro no muestra un nivel muy bajo de aceptación, su calificación final está muy por debajo del resto de repositorios en su categoría. Vale resaltar que se puede tener acceso a los datos estructurados de cada repositorio a través de RSS y de Atom

4.Implementación de las Herramientas de Búsqueda

Propósito

Instalar y configurar las herramientas de búsqueda para recuperar material educativo.

Introducción

La implementación de las herramientas ha sido realizada considerando lo estudiado y definido previamente, así en la fase 1: “Estudio de Sistemas de Búsqueda para Recursos Educativos Abiertos”, se presentaron varias herramientas para la implementación, de las cuales se seleccionaron dos (Nutch y Regain),

Estudiadas las herramientas para la recuperación de información; las mismas que se ajustan a la arquitectura general de un SRI, se deberá integrar todo este conocimiento con el fin de poder realizar las búsquedas a partir del material educativo residente en los distintos repositorios, los cuales fueron seleccionados de acuerdo a los criterios mostrados en la fase 3: “Análisis y Selección de Fuentes de Material Educativo para la Búsqueda”. En este capítulo también se aborda la instalación de las herramientas seleccionadas así como su integración con los repositorios seleccionados.

4.1 Esquema de Recuperación para los Repositorios de Material Educativo Seleccionados

Como se presentó en el capítulo 1, sección 1.2.6, un SRI tiene algunos componentes básicos, entre otros: la interfaz de usuario, un motor de búsqueda, un indexador, un índice, y un crawler.

Las dos herramientas consideradas en este proyecto, se ajustan al esquema de recuperación de información tradicional, Sección 1.2.6; sin embargo, para asegurar la presentación de material educativo de calidad, en este trabajo, en lugar de que el crawler realice sus búsquedas en toda la web como los buscadores tradicionales, las herramientas realizarán la búsqueda sobre los recursos de los repositorios OERs seleccionados previamente (Figura 4.1))

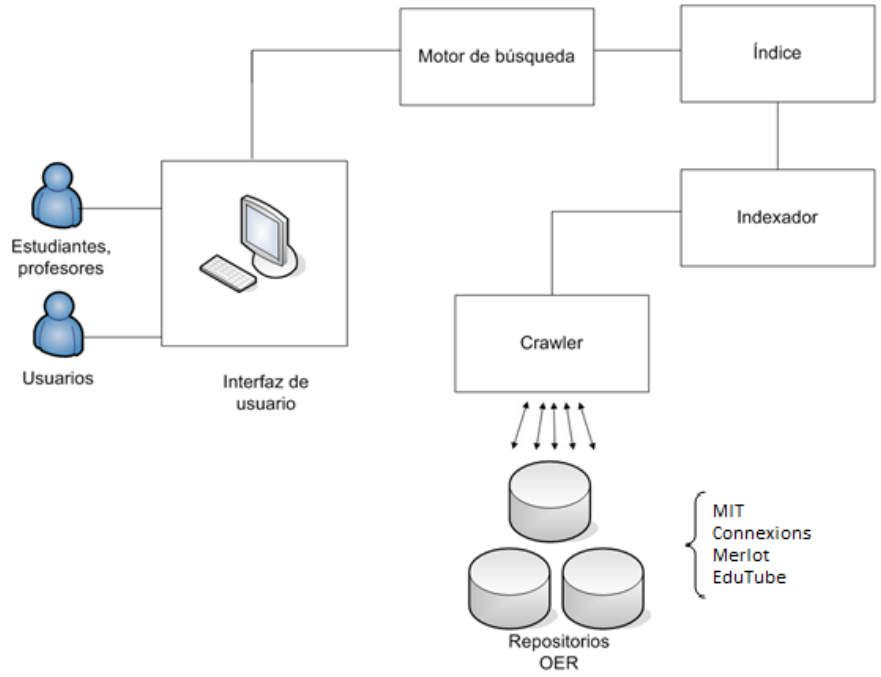

Figura 4.1 Esquema de recuperación para recuperar OERs desde distintos repositorios.

4.2 Implementación de Herramientas

A continuación se describen ciertos aspectos de la instalación de las herramientas seleccionadas, Nutch y Regain.

4.2.1 Características Hardware y Software

Características del equipo donde se realizó la instalación

Las características del equipo tanto de hardware como de software donde se realizaron las instalaciones se especifican a continuación:

Tabla 4-1 Características del equipo.

Características de Hardware	
Marca del equipo	Dell
Modelo	Inspiron 1525
Procesador	Intel Core 2 Duo 2.0 GHz
Disco duro	230 GB
Memoria RAM	3 GB
Características de Software	
Sistema operativo	Microsoft Windows 7 Ultimate 32 bits
Navegador	Mozilla Firefox 3.6.8
Java	1.6.0_21

Características del software utilizado

El software utilizado para la instalación de las herramientas es el siguiente

Tabla 4-2 Características del software utilizado.

Regain	
Regain	1.6.6
Java	JDK 1.6.0_21
Nutch	
Nutch	1.1
Java	JDK 1.6.0_21
Apache Tomcat	6.0.29
Cygwin	1.7.5

4.2.2 Instalación y Configuración de las Herramientas

La instalación de Apache Nutch 1.2 y de Regain se explica detalladamente en los **Anexos 2 y 3** respectivamente; la herramienta Regain es muy fácil de instalar y muy intuitiva para usar, siendo Nutch más compleja de configurar, por lo que en esta sección se detallará los procesos seguidos para lograr una configuración de esta herramienta:

4.2.2.1 Configuración de Nutch

Configuración de plugins de Nutch 1.1

Para empezar con la configuración, vale recalcar que por defecto Nutch 1.1 solo realiza búsquedas basadas en Lucene; es decir, analizando el texto plano solamente. Como se menciona en el Anexo 8 sección 8.1.3.4 (Plugins disponibles con Nutch), Nutch dispone de un

conjunto de plugins los cuales se pueden incluir para realizar una búsqueda mucho más compleja, organizar y extraer el contenido de una forma mucho más eficaz y desde diferentes tipos de recursos como por ejemplo java script, RSS, pdf, documentos de Microsoft office y open office, etc.

Para la utilización de cualquiera de estos plugins, se debe incluir en el archivo llamado 'nutch-site.xml' ubicado en la carpeta 'conf' del directorio donde se instaló nutch la propiedad 'plugin.includes', como se muestra en la figura siguiente:

```
<property>
 <name>plugin.includes</name>
   <value>
    protocol-http|urlfilter-regex|parse
   (text|html|js|tika|pdf|rss|ext|xml|feed|index-basic|query-
   (basic|site|url|summary-basic|scoring-opic|urlnormalizer-
   (pass|regex|basic)|nutch-extensionpoints|index-more|query-
   more|summary-lucene|creativecommons
  </value>
  <description>
   Por defecto Nutch incluye crawling solo para HTML y texto
   plano via HTTP.
  </description>
</property>
```

Figura 4.2 Utilización de plugins disponibles en Nutch 1.2

Configuración del archivo parse-plugin

El archivo 'parse-plugins.xml' ubicado en la carpeta 'conf' del directorio donde se instaló nutch representa un orden natural por el cual el analizador de plugins debe llamar a un determinado mimeType[33]. Para la extracción del contenido de los repositorios seleccionados se tuvo que configurar los siguientes mimeType:

Para el repositorio MIT se tuvo que configurar mimetype "application/rdf+xml":

```
<mimeType name="application/rdf+xml">
  <plugin id="parse-html" />
  <plugin id="parse-rss" />
  <plugin id="feed" />
  <plugin id="creativecommons" />
  <plugin id="tika" />
</mimeType>
```

Figura 4.3 Configuración mimeType: "rdf+xml"

[33] Una manera de describir el tipo de documento a ser transmitido (Apache_Software_Foundation, 2010).

Para los repositorios Connexions y Merlot se tuvo que configurar mimetype “xhtml+xml” respectivamente:

```
<mimeType name="application/xhtml+xml">
 <plugin id="parse-html" />
 <plugin id="feed" />
 <plugin id="creativecommons" />
 <plugin id="tika" />
</mimeType>
```

Figura 4.4 Configuración mimeType: “xhtml+xml”

Para el mimeType que ofrece EduTube ya existe un mimeType en el archivo ‘parse-plugin’ por defecto, sin embargo se le añadió el plugin de ‘creativecommons’ y ‘tika’ para una mejor extracción de los datos, así mismo se realizó con el mimeType “text/xml”:

```
<mimeType name="application/rss+xml">
 <plugin id="parse-rss" />
 <plugin id="feed" />
 <plugin id="creativecommons" />
 <plugin id="tika" />
</mimeType>
```

Figura 4.5 Configuración mimeType: “rss+xml”

```
<mimeType name="text/xml">
     <plugin id="parse-html" />
     <plugin id="parse-rss" />
     <plugin id="feed" />
     <plugin id="creativecommons" />
     <plugin id="tika" />
</mimeType>
```

Figura 4.6 Configuración mimeType: “text+xml”

Configuración del archivo ‘crawl-urlfilter’

El archivo ‘crawl-urlfilter’ ubicado en la carpeta ‘conf’ del directorio donde se instaló nutch, es donde se le indicará al crawler qué debe o no indexar, actúa como un filtro al momento de realizar la indexación, siendo las expresiones regulares “+” y ”-” utilizadas como prefijo de cada línea, indicando si algo se debe o no tomar en cuenta respectivamente. Por defecto este archivo omite los siguientes caracteres especiales: [?, *, !, @, =], siendo esto configurable; en este caso algunas páginas RSS en las cuales se va a realizar la extracción de los metadatos contienen caracteres especiales como [?=+] por lo que este parámetro se configuró de la siguiente manera:

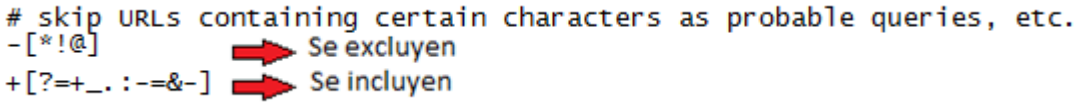

Figura 4.7 Configuración del archivo 'crawl-urlfilter'.

De igual manera en este archivo se antepone el símbolo '**#**' a cualquier línea que nos e vaya a tomar en cuenta por Nutch, en este caso se delimito una sección la cual no permite al crawler buscar enlaces que contengan el símbolo '/' más de tres veces. Esto se realizó ya que existen páginas a indexar que contienen un número mayor que tres.

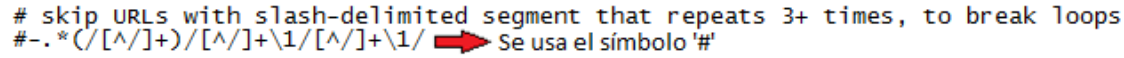

Figura 4.8 Configuración del archivo 'crawl-urlfilter' - delimitador.

Así mismo en este archivo se deben ingresar las páginas web que van o no ser indexadas dentro de un determinado dominio, las páginas a indexar se encuentran en el **ANEXO 9.7** 'Configuración del archivo 'crawl-urlfilter.txt' de Nutch'.

Finalmente, los enlaces web en los cuales se realizará la búsqueda se deben incluir en el archivo 'otech_crawl.txt'; los enlaces utilizados para la evaluación del SRI se encuentran en el **ANEXO 9.6** 'Configuración del archivo 'otech_crawl.txt' -Enlaces web de la categoría Antropología de los repositorios'.

4.3 Resultados Preliminares

Este capítulo concluye con la instalación de las herramientas antes mencionadas, las mismas que se adaptan a la arquitectura propuesta y son capaces de realizar las búsquedas desde diferentes repositorios. A continuación se muestran algunas imágenes de las herramientas ya en funcionamiento como resultados preliminares:

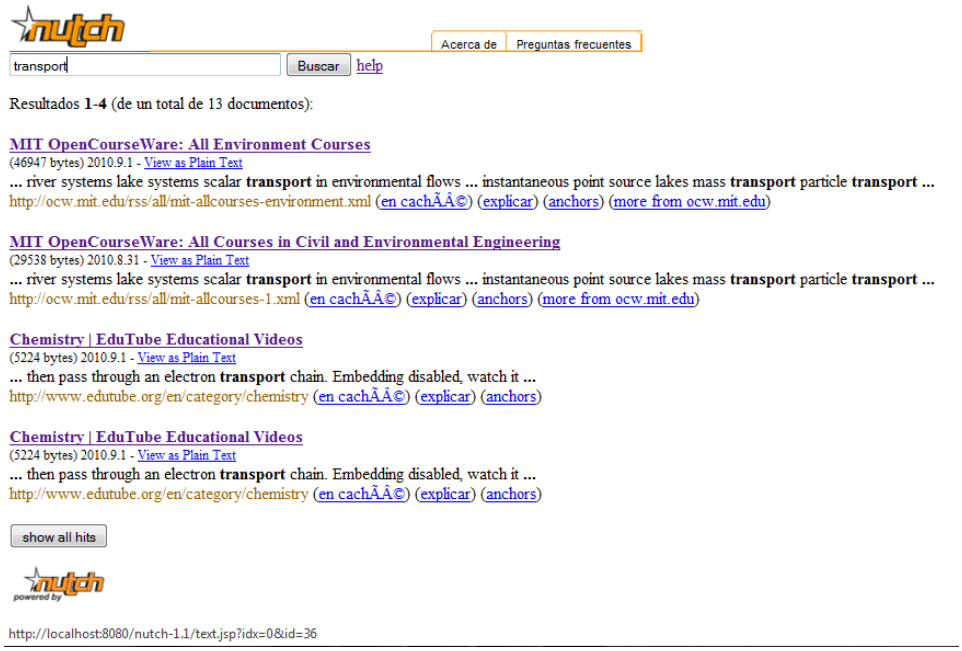

Figura 4.9 Búsqueda utilizando la herramienta Nutch.

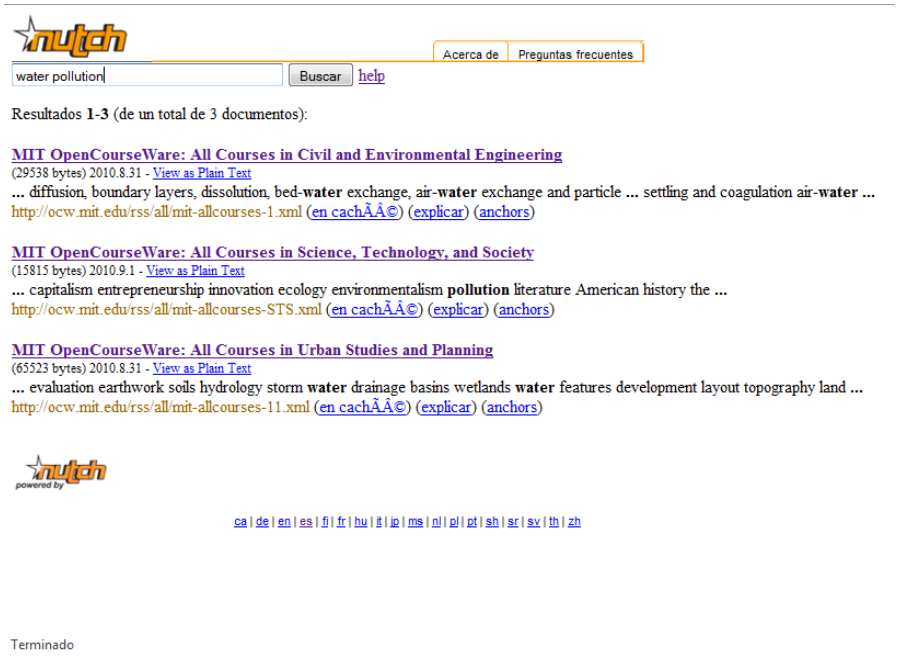

Figura 4.10 Búsqueda utilizando la herramienta Nutch (Muestra 2).

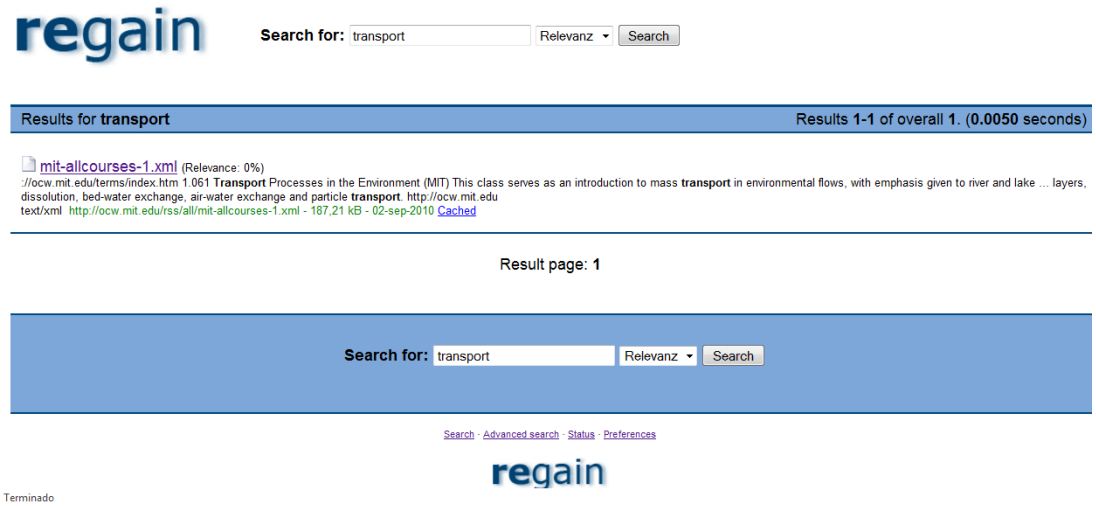

Figura 4.11 Búsqueda utilizando la herramienta Regain.

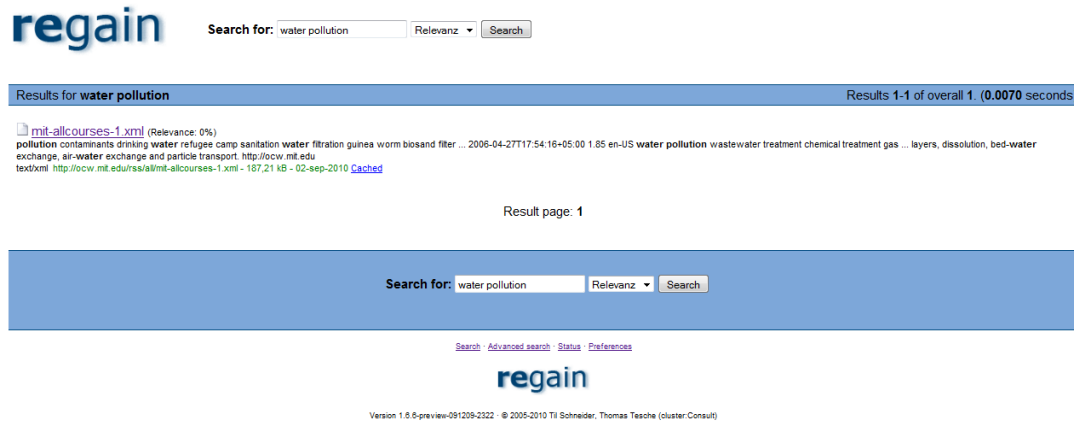

Figura 4.12 Búsqueda utilizando la herramienta Regain (Muestra 2).

4.4 Discusión

Cómo se pudo apreciar a lo largo de este capítulo, la herramienta Regain resultó muy sencilla de implementar, esto se facilita aún más gracias a su asistente de instalación, por lo que no representó ningún inconveniente; por el contrario, la herramienta Nutch presentó muchos obstáculos al momento de implementarla; todos estos obstáculos fueron oportunamente superados y cómo resultado de ello en el **Anexo 2** se presenta la instalación y configuración paso a paso de Nutch y que junto con la sección 4.3.2.1: "Configuración de nutch" de este capítulo complementa una correcta configuración de la misma.

En cuanto a las características de hardware, ambas herramientas respondieron satisfactoriamente y sin ningún inconveniente. Al ser un SRI que realiza sus búsquedas en cuatro repositorios determinados, la cantidad de información a indexar es relativamente poca frente a un buscador web tradicional, por lo que el tamaño del índice no representa mayor inconveniente de almacenamiento.

5.Comparación de las Herramientas de Búsqueda para Material Educativo

Propósito

Evaluar y comparar las herramientas de recuperación para recursos educativos abiertos.

Introducción

Para evaluar el rendimiento de los sistemas de búsqueda, varias medidas han sido propuestas, sin embargo, dos de esas medidas son ampliamente utilizadas: la exhaustividad y la precisión, conceptos ya analizados en el capítulo 2. En ambos casos, la medidas se basan en la relevancia de los documentos recuperados, es decir, en qué tanto se ha satisfecho la necesidad de información de los usuarios que hacen la consulta (Serna et al., 2004).

Aunque la relevancia es un criterio subjetivo, debido a que diferentes personas asignarían diferentes valores de relevancia a un documento, siempre se toma en cuenta en cualquier método de evaluación de los SRI (Serna et al., 2004).

En este trabajo, se realizó una medición de la precisión y exhaustividad considerando el criterio de relevancia del autor, de acuerdo a consultas establecidas dentro del área de conocimiento de Antropología.

De los 4 repositorios de material educativo seleccionados, se realizó un análisis de la cantidad de material educativo disponible por cada área de conocimiento, resultando ser la de Antropología la que tenía menor número, por tanto, se seleccionó esta área porque la evaluación de cada resultado fue realizado de forma manual e individual (por cada resultado).

Luego de realizar la evaluación del rendimiento de las herramientas Nutch y Regain, en este capítulo también se realiza la evaluación considerando los criterios definidos en la sección 2.2.6, los cuales intentan medir el nivel de aceptación de usuarios finales.

En la sección 5.1 de este capítulo se realiza un estudio para validar que los requerimientos del sistema de búsqueda estudiados en el capítulo 3 hayan sido los correctos; y en la sección 5.2 se realiza la implementación de la metodología seguida, donde finalmente se concluye con un análisis de los resultados obtenidos.

5.1 Evaluación de Criterios a Nivel de Usuario

En el capítulo 2. 'Evaluación del sistema de búsqueda', se realizó un estudio para determinar si los dos sistemas de búsqueda seleccionados cumplen con los requerimientos que un sistema de recuperación deberá tener. En base a la [Tabla 2-1] de dicho capítulo se presenta a continuación una tabla, en la cual se indican las respectivas calificaciones que fueron proporcionadas por el autor de este proyecto a cada una de estas herramientas, para luego realizar una comparación en base a una encuesta realizada al usuario final y así poder determinar si es que realmente se cumplió con estos pre-requerimientos.

Tabla 5-1 Calificación inicial del SRI.

		Calificación (Personal)					
		Nutch			Regain		
	Requerimiento	Baja	Media	Alta	Baja	Media	Alta
Requerimientos	Amigabilidad			√			√
	Informatividad		√			√	
	Visualización de recuperación		√			√	
	Búsqueda avanzada	√				√	
	Eficacia			√			√
	Eficiencia			√			√
	Integridad			√			√
	Disponibilidad			√			√
	Confiabilidad		√			√	
	exhaustividad		√			√	
	La precisión		√		√		
		SI		NO	SI		NO
Tecnologías semánticas	Metadatos	√					√
	Rss, Atom, Feeds	√					√
	RDF	√					√
	RDFa			√			√
	Microformatos	√					√

Cómo se puede observar en la [Tabla 5-1], ambas herramientas, tanto Regain como Nutch, presentan en su mayor parte características similares en cuanto a sus requerimientos se refiere. La diferencia fundamental entre ambas radica en las tecnologías semánticas sobre las cuales pueden trabajar; Regain no trabaja sobre los datos estructurados, solo realiza sus búsquedas en texto plano. Por el contrario, Nutch permite realizar búsquedas sobre la mayoría de estas tecnologías.

Para comprobar que esta calificación inicial haya sido correcta y evitar cualquier error sistemático que conduzca a una estimación incorrecta de la calificación otorgada inicialmente debido a una calificación individual (propia); se ha realizado una encuesta a diez usuarios finales a fin de verificar si es que verdaderamente se cumple la calificación otorgada a los requerimientos inicialmente. En base a los resultados obtenidos de la encuesta se obtuvo como resultado la calificación mostrada en la tabla siguiente; siendo los valores resaltados la calificación más alta por parte de los usuarios, teniendo en cuenta el número de votos por cada calificación.

Tabla 5-2 Calificación de los usuarios para el SRI I

	Nutch			Regain		
Requerimiento	**Baja**	**Media**	**Alta**	**Baja**	**Media**	**Alta**
Amigabilidad	0	1	**9**	0	1	**9**
Informatividad	2	2	**6**	3	**6**	1
Visualización de recuperación	0	4	**6**	1	**4**	**4**
Búsqueda avanzada	**10**	0	0	0	**8**	2
Eficacia	0	0	**10**	0	0	**10**
Eficiencia	0	1	**9**	0	2	**8**
Integridad	1	1	**8**	**5**	**5**	0
Disponibilidad	0	1	**9**	0	1	**9**
Confiabilidad	1	0	**9**	1	**7**	2
Exhaustividad	2	0	**8**	0	**7**	3
La precisión	1	3	**6**	2	**8**	0

Para realizar una comparación posterior se ha transcrito la **[Tabla 5-2]** como se muestra a continuación:

Tabla 5-3 Calificación de los usuarios para el SRI II

	Nutch			Regain		
Requerimiento	**Baja**	**Media**	**Alta**	**Baja**	**Media**	**Alta**
Amigabilidad			O			O
Informatividad			O		O	
Visualización de recuperación			O		O	O
Búsqueda avanzada	O				O	
Eficacia			O			O
Eficiencia			O			O
Integridad			O	O	O	
Disponibilidad			O			O
Confiabilidad			O		O	
exhaustividad			O		O	
La precisión			O		O	

En la [Tabla 5-3] se puede observar la calificación que obtuvieron ambos sistemas de búsqueda por parte de los usuarios finales. Cabe recalcar que en algunos requerimientos se puede apreciar el mismo puntaje para ambas calificaciones.

Una vez obtenidos los resultados por parte de los usuarios, se ha realizado una comparación de los resultados obtenidos en las tablas [5-1] y [5-3], se han unificado ambos resultados en una misma tabla la cual va a constar de la siguiente nomenclatura:

X = Resultados que coinciden

O = Calificación por parte de los usuarios finales

√ = calificación otorgada inicialmente (propia)

Tabla 5-4 Comparación de los resultados obtenidos al evaluar el SRI

	Nutch			Regain		
Requerimiento	**Baja**	**Media**	**Alta**	**Baja**	**Media**	**Alta**
Amigabilidad			X			X
Informatividad		√	O		X	
Visualización de recuperación		√	O		X	O
Búsqueda avanzada	X				X	
Eficacia			X			X
Eficiencia			X			X
Integridad			X	O	O	√
Disponibilidad			X			X
Confiabilidad		√	O		X	
exhaustividad		√	O		X	
La precisión		√	O	√	O	

Para entender mejor los resultados obtenidos y con el fin de determinar si la calificación otorgada inicialmente se cumple o no, se analizaran estos resultados en base a las siguientes figuras, las cuales sintetizan los resultados obtenidos de la comparación realizada en la [Tabla 5-4]:

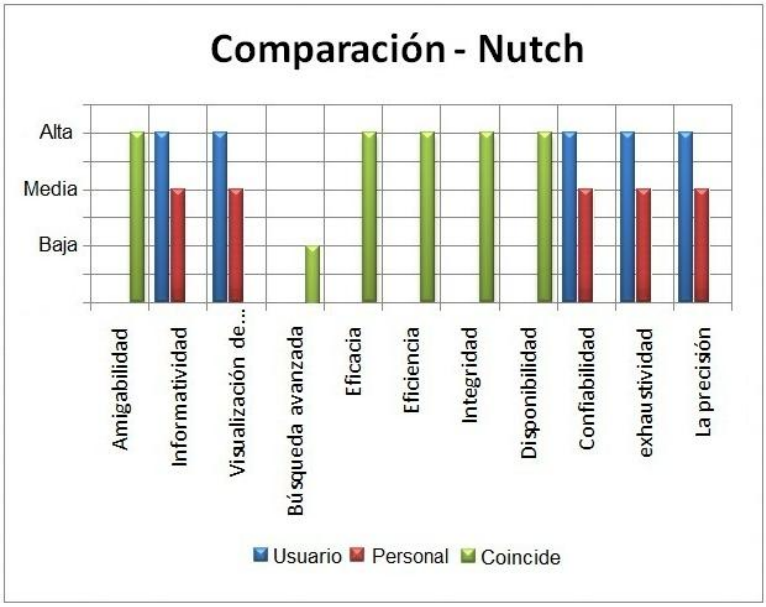

Figura 5.1 Comparación de calificaciones para Nutch

Primero, para la herramienta Nutch se puede observar en la **Figura 5.1** que, de los 11 requerimientos, 6 de ellos coinciden con la calificación proporcionada por el autor del proyecto, y de los 5 requerimientos restantes que son: informatividad, visualización de recuperación, confiabilidad, exhaustividad y precisión; el usuario tuvo en todos estos casos una mayor aceptación de lo planteado inicialmente, dando una calificación alta a todos estos requerimientos.

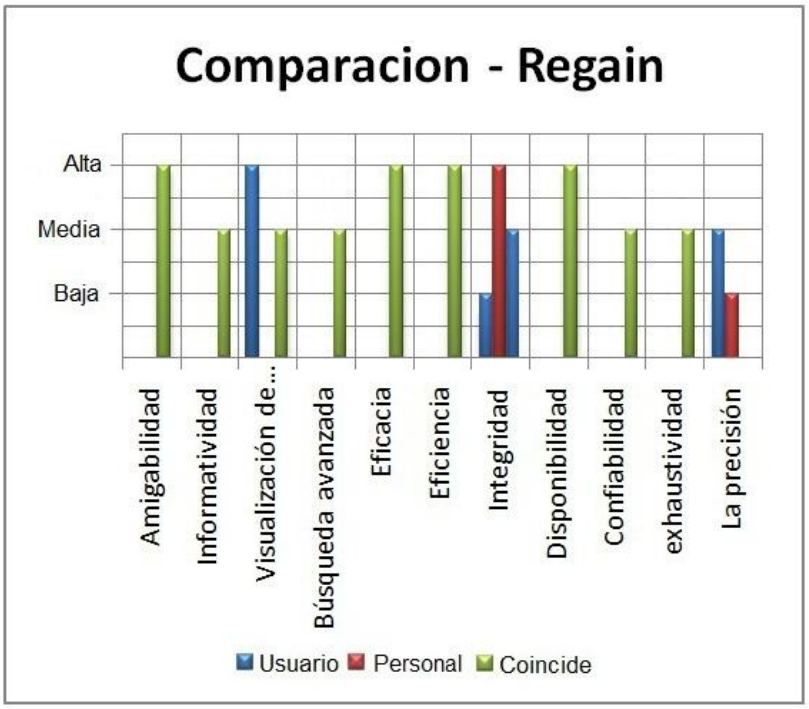

Figura 5.2 Comparación de calificaciones para Regain

Así mismo se puede observar en la Figura 5.2, que de los 11 requerimientos, 9 de ellos coinciden con la calificación proporcionada por el autor del proyecto, y los dos restantes que son: integridad y precisión demostraron tener una calificación inferior a la planteada inicialmente. En base a estos análisis se puede concluir que, las calificaciones dadas inicialmente a ambos SRI son en su mayor parte correctas; y como se puede apreciar en el análisis realizado, la herramienta Nutch tiene una calificación Alta en comparación a Regain.

En el capítulo 1 se explicó que Nutch es una herramienta que nos ofrece varias opciones al momento de realizar una búsqueda, mientras que Regain ofrece búsquedas basadas solo en texto plano, pero será tomado en cuenta en este estudio de tesis para determinar el rendimiento que se obtiene al realizar una búsqueda basada en datos estructurados como por ejemplo RDF vs una búsqueda convencional. Siendo esta la principal diferencia entre ambas herramientas, y es por esta razón que la herramienta Nutch fue seleccionada para realizar la búsqueda de OER; al poder realizar búsquedas y tomando los datos estructurados disponibles para ella, la precisión y la exhaustividad del SRI aumenta; esto será estudiado en la siguiente sección de este capítulo.

5.2 Evaluación del Rendimiento

Cómo se observa en el **Anexo 4**: “Disponibilidad de cursos en los repositorios seleccionados”, los cuatro repositorios tienen diferentes categorías, de las cuales las que son comunes entre ellos o en su gran mayoría son las siguientes:

- Art
- Historia
- Music
- Chemistry
- Mathematics
- Physics
- Technology
- Anthropology
- Politics
- Health
- Econimics
- Engineering
- Philosophy

Con fines de pruebas, se tomará una categoría de entre las presentadas, a continuación se ilustran estas categorías con el número correspondiente de recursos disponibles en cada una de ellas:

Tabla 5-5 Número de recursos por categoría y repositorio.

		Repositorios				
		Edutube	**Mit**	**Merlot**	**Connexions**	**Total**
Categorías Comunes	**Anthropology**	19	39	64	83	205
	chemistry	23	43	720	318	1104
	Philosophy	9	115	587	408	1119
	Politics	37	158	200	951	1346
	Music	60	54	297	1234	1645
	Economics	31	85	586	1054	1756
	Health	52	61	1371	866	2350
	Art	60	70	1112	1285	2527
	Historia	50	62	969	1792	2873
	Engineering	13	709	715	1439	2876
	Physics	45	69	1813	1908	3835
	Mathematics	51	125	1523	3800	5499
	Technology	121	61	915	6239	7336

Como se puede observar en la [Tabla 5.5] la categoría 'Anthropology' presenta el menor número de elementos totales por categoría con un total de 205 elementos; un número relativamente bajo en comparación de las demás categorías las cuales van desde 1000 en adelante hasta un máximo de 7336; hay que tener en cuenta que este número máximo no representa el mayor número de elementos por categoría en todos los repositorios, puesto que este análisis se realizó en las categorías comunes entre los repositorios seleccionados.

La categoría 'Anthroplogy' ha sido tomada como base para realizar la evaluación; y aunque la tendencia al evaluar un SRI en internet es seguir el método 2 descrito en la sección 2.1.2: 'Cálculo de la precisión y de la exhaustividad'; en el cual se analizan los primeros N resultados devueltos por el SRI, esto resulta posible debido a la gran cantidad de ítems existentes; pero por la naturaleza del SRI propuesto, esto presenta dificultades, ya que al ser un SRI que solo realiza su búsqueda en determinados repositorios (concretamente cuatro), los resultados devueltos serán mínimos o inclusive no se devolverán resultados debido a la poca cantidad de ítems, posiblemente únicos en su determinada categoría. Es por esta razón que se realizará un cálculo manual de la precisión y exhaustividad en una muestra pequeña.

a) Determinación de las necesidades de información de los usuarios.

En el presente trabajo se ha tomado un número de ocho preguntas, esto debido a que solo se realizarán pruebas sobre la categoría 'Anthropology' de los cuatro repositorios seleccionados como se observa en la [Tabla 5.6].

El lenguaje utilizado para la realización de las búsquedas fue el inglés, considerando que los repositorios sobre los cuales se realizará la extracción de los datos están en dicho idioma, a continuación se elabora el enunciado de las ocho preguntas referentes al tema de antropología:

Tabla 5-6 Necesidades de información – Preguntas a realizarse.

Pregunta 1	Information about the "islas uros"
Pregunta 2	A museum online or virtual
Pregunta 3	What is a shaman, what it's the meaning of "shaman"?
Pregunta 4	The Mayan civilization
Pregunta 5	An introduction to Anthropology
Pregunta 6	Who built the pyramids?
Pregunta 7	The conquest of America
Pregunta 8	Videos about anthropology.

b) Elaboración del enunciado de búsqueda (La sintaxis de búsqueda).

En base a las preguntas planteadas en la [Tabla 5.6]: "Necesidades de información – Preguntas a realizarse."; ahora se plantean dichas preguntas en el buscador de distinta manera, tratando de que se obtenga el mayor número de resultados relevantes posibles, obteniendo como resultado las siguiente sintaxis de búsqueda:

Tabla 5-7 Elaboración de la sintaxis de búsqueda.

Pregunta 1	The islas uros
Pregunta 2	museum online
Pregunta 3	What is a shaman
Pregunta 4	mayan civilization
Pregunta 5	Introduction to Anthropology
Pregunta 6	who build the pyramids
Pregunta 7	conquest of america
Pregunta 8	video of anthropology

c) Realización de las consultas.

Cómo muestra de esto tenemos un ejemplo de la pregunta 8 planteada al buscador Nutch:

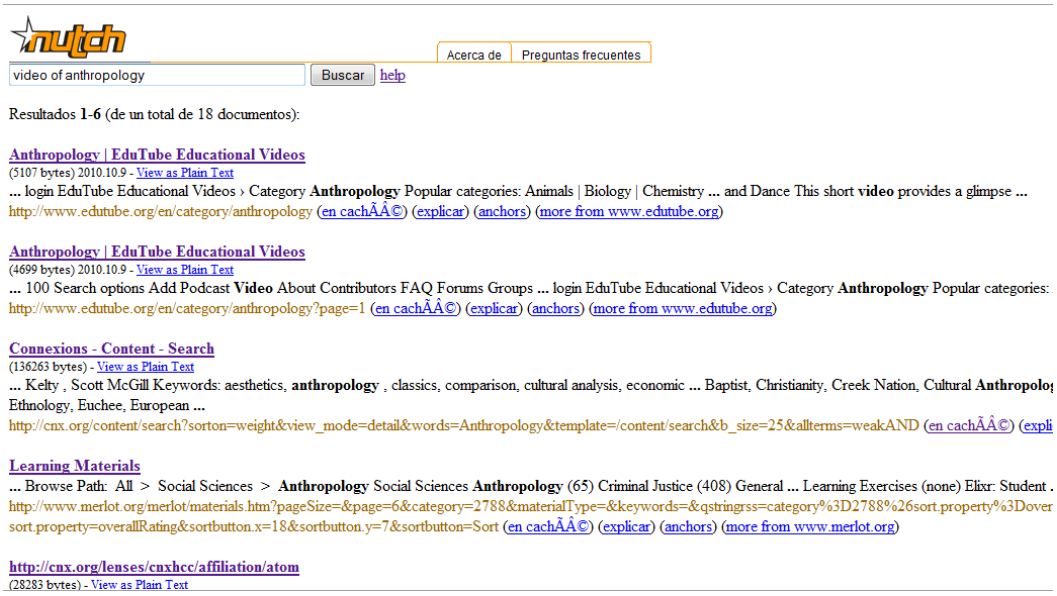

Figura 5.3 Realización de una consulta en el buscador Nutch.

Figura 5.4 Realización de una consulta en el buscador Regain.

d) Valoración de la relevancia.

De acuerdo al **Anexo 5**: "Archivos disponibles para la categoría 'Anthropology' ", se ha revisado cada documento/curso de la categoría antropología de cada uno de los repositorios seleccionados; así, se han determinado de acuerdo a la [Tabla 2.2]: "Escala de evaluación para medir la relevancia (Olvera, 2000)." qué documentos son relevantes frente a la pregunta planteada, este es un proceso tedioso ya que se realiza manualmente, pero sus resultados son realmente buenos. Frente a esto se ha realizado una matriz, en la cual se muestran qué documentos/cursos son relevantes para cada pregunta, así a cada curso le corresponde un ID único el cual nos servirá como referencia para encontrar dichos documentos (Véase anexo 5).

Tabla 5-8 Matriz - Enlaces Relevantes por pregunta.

		Repositorios				
	Búsqueda	**EduTube**	**MIT**	**Connexions**	**Merlot**	
Pregunta 1	The islas uros				MER014 MER027 MER028 MER036 MER037	ID
Pregunta 2	museum online				MER001 MER025 MER054	
Pregunta 3	What is a shaman			CON013		
Pregunta 4	mayan civilization	EDU020			MER016	
Pregunta 5	Introduction to Anthropology		MIT013	CON001	MER043 MER044 MER053	
Pregunta 6	who build the pyramids	EDU005 EDU006 EDU007				
Pregunta 7	conquest of america		MIT029			
Pregunta 8	video of anthropology	EDU001 …. EDU020		CON001	MER036 MER037 MER055	
	Total por Repositorio	24	2	3	15	

Una vez determinados los documentos/cursos que son relevantes para cada pregunta planteada, se procede a calcular para cada pregunta el número de enlaces que devuelve y cuantos documentos relevantes se recuperan en dicha pregunta; este proceso se realizó en ambos buscadores (Nutch y Regain), obteniéndose los siguientes resultados:

Tabla 5-9 Documentos relevantes recuperados – SRI NUTCH.

		Enlaces recuperados Totales	Enlaces recuperados Relevantes	Documentos recuperados Relevantes en los enlaces	Documentos relevantes en los repositorios
Pregunta 1	The islas uros	2	2	5	5
Pregunta 2	museum online	4	3	3	3
Pregunta 3	What is a shaman	0	0	0	1
Pregunta 4	mayan civilization	3	1	1	2
Pregunta 5	Introduction to Anthropology	6	4	5	5
Pregunta 6	who build the pyramids	2	1	3	3
Pregunta 7	conquest of america	0	0	0	1
Pregunta 8	video of anthropology	18	16	24	24

Tabla 5-10 Documentos relevantes recuperados – SRI REGAIN.

		Enlaces recuperados Totales	Enlaces recuperados Relevantes	Documentos recuperados Relevantes en los enlaces	Documentos relevantes en los repositorios
Pregunta 1	The islas uros	0	0	0	5
Pregunta 2	museum online	3	0	0	3
Pregunta 3	What is a shaman	25	1	1	1
Pregunta 4	mayan civilization	21	1	1	2
Pregunta 5	Introduction to Anthropology	148	2	2	5
Pregunta 6	who build the pyramids	12	3	3	3
Pregunta 7	conquest of america	50	1	1	1
Pregunta 8	video of anthropology	112	21	21	24

e) Análisis de los resultados.

Como se ha indicado, las medidas utilizadas para analizar los resultados son la precisión y exhaustividad de los resultados obtenidos en las consultas; el cálculo de estas medidas resulta muy simple en este caso en particular debido a que se realizó un proceso manual sobre una categoría específica y se conoce de antemano el número total de ítems relevantes frente a determinada pregunta.

Como se mencionó en el capítulo II sección 2.1.2 ‘Cálculo de la precisión y de la exhaustividad’; estas medidas se calculan de la siguiente manera:

Para calcular la precisión se divide el total de documentos relevantes recuperados entre el total de documentos recuperados. Pero debido a la naturaleza de los buscadores Nutch y Regain que en un solo enlace se pueden encontrar dos documentos relevantes, se tomará en cuenta la precisión de los enlaces. Por ejemplo para la pregunta 2 de la [Tabla 5-9], el SRI contiene 3 enlaces relevantes para dicha pregunta; y el SRI obtiene 4 enlaces recuperados de los cuales 3 son relevantes; entonces la precisión del sistema a nivel de enlaces es de 3/4, es decir un 75%.

La exhaustividad corresponde al cociente entre el número de documentos relevantes recuperados y el total relevantes existentes en la colección. Por ejemplo para la Pregunta 4 de la [Tabla 5-8], en los repositorios existen 2 documentos relevantes para dicha consulta y el sistema de búsqueda obtuvo 1 documento relevante correspondiente a dicha pregunta; la exhaustividad será 1/2, es decir el 50%.

Así, al realizar estos cálculos sobre los resultados obtenidos en las [Tablas 5-9 y 5-10] se obtiene:

Tabla 5-11 Cálculo de precisión y exhaustividad de NUTCH.

	Enlaces recuperados Totales	**Enlaces recuperados Relevantes**	**Documentos Relevantes Recuperados**	**Documentos relevantes en los repositorios**	**Precisión (%)**	**Exhaustividad (%)**
Pregunta 1	2	2	5	5	100	100
Pregunta 2	4	3	3	3	75	100
Pregunta 3	0	0	0	1	0	0
Pregunta 4	3	1	1	2	33,33	50
Pregunta 5	6	4	5	5	66,67	100
Pregunta 6	2	1	3	3	50	100
Pregunta 7	0	0	0	1	0	0
Pregunta 8	18	16	25	25	88,89	100
				Media →	51,74	68,75

Tabla 5-12 Cálculo de la precisión y exhaustividad de REGAIN.

	Enlaces recuperados Totales	Enlaces recuperados Relevantes	Documentos Relevantes Recuperados	Documentos relevantes en los repositorios	Precisión (%)	Exhaustividad (%)
Pregunta 1	0	0	0	5	0	0
Pregunta 2	3	0	0	3	0	0
Pregunta 3	25	1	1	1	4	100
Pregunta 4	21	1	1	2	4,76	50
Pregunta 5	80	2	2	5	2,5	40
Pregunta 6	20	3	3	3	15	100
Pregunta 7	50	1	1	1	2	100
Pregunta 8	112	21	21	25	18,75	84
				Media →	5,88	59,25

Al analizar las tablas anteriores se observa que a Nutch le corresponde una precisión de un 51,74% frente a un 5,8% de Regain; así mismo la exhaustividad para Nutch es de un 68,75% frente Regain que obtuvo un 59,25 %. Esto se resume en la siguiente tabla:

Tabla 5-13 Valores de precisión y exhaustividad obtenidos en la evaluación de los sistemas de búsqueda.

	Nutch	Regain
Precisión	51,74 %	5,88%
Exhaustividad	68,75 %	59,25%

Con estas pruebas realizadas, se puede concluir diciendo que Nutch resulta mucho más efectivo que Regain a la hora de responder a una pregunta planteada, su grado de precisión es mucho mayor; es decir, la mayor parte de documentos recuperados que resultan de una pregunta son relevantes frente al tema; por otro lado, la exhaustividad tanto en Nutch como en Regain resulta en un valor próximo, siendo la diferencia entre ambos de un 9,5 %. La exhaustividad mide la porción de documentos relevantes que son recuperados, y aunque Regain no resulta preciso al momento de responder una consulta, arroja un gran número de resultados en cada pregunta, y dentro de este número de resultados obtenidos se encuentran algunos de los documentos relevantes conocidos previamente; pero esto se conoce ya que se realizó un análisis manual de los documentos y se analizaron todos los enlaces arrojados por Nutch y Regain; es por esto que el valor de exhaustividad de Regain se asemeja al de Nutch;

pero en realidad, un usuario no va a buscar en todos los enlaces obtenidos de una consulta, generalmente revisa los primeros resultados arrojados por el buscador.

Como se puede observar en la [Tabla 5-8] el repositorio que más recursos aporto en todas las preguntas en conjunto fue MERLOT, no se tomó en cuenta a EduTube puesto que la pregunta número 7 tiene por objetivo obtener videos acerca de antropología, y siendo EduTube un repositorio propio de videos obtendremos como resultado todos o la mayor parte de recursos de este repositorio.

Acceder a las fuentes RSS y Atom de los distintos repositorios para realizar la búsqueda mejora notablemente los resultados obtenidos; los datos estructurados permiten obtener respuestas más precisas y concretas acerca del tema, la herramienta Nutch puede permite acceder a estas fuentes a diferencia de Regain que cómo ya se mencionó anteriormente solo realiza sus búsquedas en texto plano, ya sea HTML, documentos MS Word o PDFs.

6.Conclusiones y Recomendaciones

6.1 Conclusiones

- El obstáculo para los usuarios que buscan recursos educativos en internet no es la falta de materiales educativos, sino la dificultad para encontrar dichos recursos, que usualmente son buscados a través de los motores de búsqueda.

- Un sistema de búsqueda específico para OERs resulta muy útil a la hora de querer encontrar dichos recursos, ya que no es lo mismo realizar una consulta en un buscador web tradicional que en un buscador dedicado a OERs, pues con este último se mejora la precisión; y aunque los metadatos adecuados se usen para asegurar que un buscador llegue una página determinada, la cantidad de recursos en internet es demasiado grande, por lo que será muy poco probable encontrar un determinado recurso entre tanta información; por lo que al buscar en un SRI específico, se tiene la certeza de que todos los resultados obtenidos son OERs.

- Al tener un sistema de búsqueda que realice sus búsquedas en repositorios OER, el tema de licenciamiento para el SRI propuesto no representa ningún problema, ya que todos los materiales disponibles en los repositorios seleccionados han sido analizados previamente por cada repositorio OER en particular y disponen de una licencia libre.

- La herramienta Nutch demostró ser mucho más precisa que Regain, esto gracias al uso de datos estructurados, ya que todo elemento estructurado representa una ventaja frente a uno que no lo es, proporcionan respuestas de búsqueda más específicas y concretas y mejoran la exhibición de los materiales apropiados y deseados.

- Al contar con estándares para la publicación de los datos estructurados, la búsqueda en un SRI se puede realizar sin ningún problema sobre repositorios heterogéneos (wikis, blogs, repositorios audio visuales, etc.) siempre y cuando el SRI soporte dichos estándares.

- De los repositorios seleccionados en esta memoria de tesis, y en base al estudio realizado en el capítulo 4, se pudo determinar que el Repositorio MIT (OPEN COURSEWARE CONSORTIUM) cumple con todos los criterios que un repositorio de materiales educativos debería tener, seguido por el repositorio MERLOT. Así mismo se determino que tanto los repositorios EduTube como Uchannel, ambos con contenido audiovisual, presentan una calificación muy baja frente a los demás repositorios de propósito general.

- En base a la evaluación realizada a los usuarios, se pudo determinar que las calificaciones dadas inicialmente a ambos SRI son en su mayor parte correctas; y como se puede apreciar en el análisis realizado en el capítulo 6, la herramienta Nutch tiene una calificación Alta en comparación a Regain.

- Con los resultados obtenidos en la evaluación del capítulo 6, se pudo determinar que el nivel de precisión del SRI Nutch fue de un 51,74 % frente a un nivel de precisión de un 5,88 % correspondiente al SRI Regain. Y aunque con poco más del 50% de precisión; el SRI Nutch utilizando datos estructurados, presenta una notable mejoría frente al SRI Regain el cual solo realiza sus búsquedas en texto plano.

- Con respecto a la exhaustividad de ambos SRI, Nutch tiene una exhaustividad de 68,75% frente a una exhaustividad de 59,25% de Regain; y aunque estos valores se asemejen, la diferencia entre el nivel de precisión de ambos buscadores será el factor determinante al momento de obtener los resultados esperados para una determinada consulta.

6.2 Recomendaciones

- Para encontrar recursos educativos abiertos, lo más factible es realizar las búsquedas en herramientas propias y especializadas para este propósito.

- Antes de querer reutilizar cualquier recurso educativo abierto, se debe verificar que tenga una licencia de uso libre.

- Para la creación de cualquier material educativo, sitio web o repositorio; se debe tener presente el uso de datos estructurados.

- Al momento de realizar cambios en los archivos de configuración de Nutch, se deben guardar los archivos con una extensión *.UTF-8, ya que el bloc de notas de Windows con el que se abren estos archivos por defecto los guarda con una extensión *.ANSI lo cual genera un problema al ejecutar el crawler.

- Si se desea obtener los datos desde los recursos RSS o Atom, ver si se ofrece este servicio para todos los contenidos del repositorio o si solo en los nuevos materiales disponibles.

- Si se está trabajando sobre el sistema operativo Windows 7, el Apache Tomcat debe ser ejecutado como administrador.

7.Bibliografía

Amat, C. B. (2005). Rendimiento de 8 sistemas de recuperación de información del espacio web español.

Apache Software Foundation. (2010). Glossary.

Asensi-Artiga, V. (1998). Propuesta de un modelo de interfaz genérica para sistemas de recuperación de información.

Atom Enabled Alliance. (2004). What is Atom?

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval.* Addison Wesley.

Baker, J. (2008). *Introducción a los recursos educativos abiertos.*

Beck, S. (1997). Why It's a Good Idea to Evaluate Web Sources. *Susan E. Beck* .

Bekkers, T. (2008). Open Educational Resources: Introduction Booklet and Webinar.

Benavides, D. K. (2009). Sistema de Recuperación de Información (SRI).

BiD. (2008). El paisaje de los repositorios institucionales open access en España.

Blair, D. (1990). *Language and Representation in Information Retrieval.* ACM.

ccLearn. (2009). Enhanced Search for Educational Resources— A Perspective and a Prototype from ccLearn. *ccLearn* .

Cid, M. C. (2005). Asignación de metadatos. *SISTEMAS AVANZADOS DE RECUPERACIÓN DE INFORMACIÓN* .

Cooper, W. S. (1973). On selecting a Measure of Retrieval Effectiveness. *American Society for Information Science* .

Córcoles, C., Hornung-Prähauser, V., Kalz, M., Minguillón, J., Naust-Schulz, V., Schaffert, S., y otros. (2007a). *Tutorial: BUSCAR Y ENCONTRAR REA (OER).*

Córcoles, C., Hornung-Prähauser, V., Kalz, M., Minguillón, J., Naust-Schulz, V., Schaffert, S., y otros. (2007b). *Tutorial: COMPARTIR REA (OER): publicación y reutilización.*

CreativeCommons. (2009). DiscoverEd FAQ. *Creative Commons* .

Cursada. (2009). Introducción a Recuperación de Información (RI).

Cutting, D. (2004). Nutch, Open-Source Web Search.

Delbru, R. (2009). SIREn.

Delbru, R. (2009). SIREn Presentation.

Donald W. Craik. (2010). Evaluation of Internet Searching and Search Engines. *Donald W. Craik Engineering Library* .

EDUTEKA. (2007). RECURSOS EDUCATIVOS ABIERTOS (REA).

Felquer, Bazan, L., & del, I. O. (2001). Rendimiento de los sistemas de recuperación de información en la Web: evaluación de los servicios de búsqueda (search engines) Google y Altavista según consultas de los Usuarios.

Fernández, R. C. (2008). Representación del Conocimiento. Web Semántica. *Universidad Carlos III de Madrid* .

Ferran, N., Pascual, M., Córcoles, C., & Minguillón, J. (2009). El software social como catalizador de las prácticas y recursos educativos abiertos. *OLCOS* .

Frakes, W. B., & Baeza-Yates, R. (1992). *Information Retrieval Data Structures & Algorithms.* Prentice-Hall.

Garcia, C. (2010). Primeros pasos con Nutch. *Otech* .

GNU. (2007). GNU General Public License.

González, J. (2008). Una mirada al mundo de los microformatos.

Graf, A. (2007). RDFA VS. MICROFORMATS. *DERI* .

Hatcher, E., Gospodnetic, O., & McCandless, M. (2008). *Lucene in Action.* Manning Publications.

Hernández, T., & Méndez, E. (2009). SISTEMAS DE BÚSQUEDA Y RECUPERACIÓN EN INTERNET (I): Directorios o índices.

Hewlett. (2005). Open Educational Resources Initiative. *The William and Flora Hewlett Foundation* .

Kowalski, G. (1997). *INFORMATION RETRIEVAL SYSTEMS Theory and Implementation.* Kluwer Academic.

Lamarca, M. J. (2009). Hipertexto, el nuevo concepto de documento en la cultura de la imagen.

López, A. G. (2006). *Modelos de Sistemas de Recuperación de Información Documental Basados en Información Lingüística Difusa.* UNIVERSIDAD DE GRANADA.

Lucene. (2010). Apache Lucene - Features. *Apache Lucene - Features* .

Lucene. (2010). Welcome to Solr.

Lucene Wiki. (2010). LuceneFAQ.

Manning, C. D., Raghavan, P., & Schütze, H. (2009). *An Introduction to Information Retrieval.* Cambridge University.

Martínez, F. J. (2002). *La recuperación y los sistemas de recuperación de información.* Universidad de Murcia.

Martínez, F. J. (2002). *Propuesta y desarrollo de un modelo para la evaluación de la recuperación de información en internet.* Universidad de Murcia.

Martínez, F. J. (2009). SRI: Los Sistemas de Recuperación de Información. (I).

Mel'nikov, Melikyan, & Maksimov. (2008). Characteristics of Information Retrieval Systems on the Internet: Theoretical and Practical Aspects.

Microformats. (2010). About Microformats. *About Microformats* .

Monge, S., & Ovelar, R. (2007). Repositorio 2.0: Dinámicas sociales para favorecer el desarrollo de comunidad en torno a un repositorio de contenidos educativos digitales.

Mortera, F. J., & Escamilla, J. G. (2008). LA INICIATIVA KNOWLEDGE HUB: UN APORTE DEL TECNOLÓGICO DE MONTERREY AL MUNDO. *Tecnológico de Monterrey* .

Nottingham, M., & Sayre, R. (2005). The Atom Syndication Format.

Nutch. (2010). Nutch.

NutchWiki. (2010). Nutch Wiki.

Olvera, M. D. (2000). Rendimiento de los sistemas de recuperación de información en la World Wide Web: revisión metodológica. *Revista Española de Documentación Científica* .

Pastor, J. A., & Artiga, V. A. (2009). Un modelo para la Evaluación de Interfaces en Sistemas de Recuperación de Información. *Universidad de Murcia* .

Pinto, M. (2009). BÚSQUEDA Y RECUPERACIÓN DE INFORMACIÓN.

Pinto, M. (2008). Evaluación de la cálida de recursos electrónicos educativos para el aprendizaje significativo.

Pollock, A., & Hockley, A. (1997). What's Wrong with Internet Searching.

Ramírez, K. D. (2009). Sistema de Recuperación de Información (SRI).

Ramos, J. P., & Hernández, G. A. (2008). Indización y Búsqueda a través de Lucene.

RDF Working Group. (2004). RDF.

RSS Advisory Board. (2009). RSS 2.0 Specification.

RSS-DEV Working Group. (s.f.). RDF Site Summary (RSS) 1.0.

Salazar, H. J., & Pinto, D. E. (2003). Recuperación de Información.

Salton, G., & McGill, M. J. (1986). *Introduction to Modern Information Retrieval.* McGraw-Hill, Inc. New York, NY, USA.

Santesteban, C. M. (2001). Acceso y recuperación de información en la World Wide Web. *Universidad Nacional de Mar de Plata* .

Schneider, T., & Tesche, T. (2010). Regain.

Schneider, T., & Tesche, T. (2009). Regain manual.

Seeley, Y. (2006). Apache SOLR.

Serna, N. L., Román, U., Osorio, N., Benito, O., Espezúa, J., & Vega, H. (2004). ESTUDIO Y EVALUACIÓN DE LOS SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN.

Simko, M., & Bieliková, M. (2009). Improving Search Results with Lightweight Semantic Search.

Smith, A. (1997). Testing the Surf: Criteria for Evaluating Internet Information Resources. *The Public-Access Computer Systems Review* .

Suárez, S. B. (2004). *BIBLIOTECA SEMÁNTICA DE WEBQUEST.* Universidad de Valladolid.

Tan, K. (2010). Lucene Tutorial.com. *Lucene Tutorial.com* .

Tane, J., Schmitz, C., & Stumme, G. (2004). Semantic Resource Management for the Web: An ELearning Application. *University of Karlsruhe* .

The Virtual Chase. (2008). Criteria for Quality in Information.

Tillman, E. N. (2003). Evaluating Quality on the Net.

Toscano, L. (2009). Innovación tecnológica en educación.

Tramullas, J. (2004). *Introducción a la Documática.*

UNESCO. (2002). Forum on the Impact of Open Courseware for Higher Education in Developing Countries.

Vargas, S. R. (2007). Buscador de Contenido. Lucene. *OpenCmsHispano* .

W3C. (2004). RDF Primer.

W3C. (2008). RDFa in XHTML: Syntax and Processing. *RDFa in XHTML: Syntax and Processing* .

W3C. (2010). Sobre el W3C. *Sobre el W3C* .

Wang, P. P. (2008). Latest step by Step Installation guide for dummies: Nutch 0.9.

Yuan, L., MaNeill, S., & Kraan, W. (2008). Open Educational Resources – Opportunities and Challenges for Higher Education. *JISC CETIS* .

8.Anexos

8.1 Anexo 1 - Estudio de las Herramientas Seleccionadas

Ya que tanto como Regain y Nutch utilizan la biblioteca Lucene, es pertinente realizar un estudio de Lucene para entender cómo funcionan estas dos herramientas.

8.1.1 Lucene

8.1.1.1 Conceptos Básicos

Lucene es una librería de búsqueda de texto que permite integrar una funcionalidad de búsqueda a una aplicación o a una página web; esto se hace mediante la adición de contenido a un índice de texto. La búsqueda entonces se realiza en este índice y se devuelven los resultados ordenados según la relevancia (Tan, 2010). Lucene es sólo el núcleo de un motor de búsqueda; como tal, no incluye ciertos elementos como una araña (spider) web o programas de análisis para formatos de documentos diferentes. Por lo que estos componentes y otros, deben ser agregadas por un desarrollador que usa Lucene (NutchWiki, 2010).

Lucene no toma en cuenta el origen de los datos, su formato, o incluso su idioma, siempre y cuando se pueda convertir en texto; esto inicialmente presenta un problema para la recuperación de información desde repositorios heterogéneos, pero como se verá más adelante, la herramienta Nutch supera este problema. Esto significa que se puede utilizar Lucene para indexar y buscar datos almacenados en archivos como: páginas web en servidores remotos, documentos almacenados en sistemas de archivos locales, archivos de texto simple, documentos de Microsoft Word, HTML o PDF, o cualquier otro formato desde el que se pueda extraer la información textual (NutchWiki, 2010). A continuación se explican los conceptos básicos de LLUCENE (Tan, 2010):

- **Búsqueda e indexación**

Lucene es capaz de lograr respuestas rápidas de búsqueda, ya que, en lugar de buscar directamente el texto, la búsqueda se realiza en un índice. El tipo de índice utilizado por Lucene se denomina índice invertido, ya que invierte una estructura de datos centrados en la página (páginas -> palabra) a una estructura de datos centrada en la palabra clave (palabras -> página).

- **Documentos**

En Lucene, un documento es la unidad de la búsqueda y la indexación. Un índice se compone de uno o más documentos, la indexación consiste en añadir documentos a una ‘IndexWriter’, y la búsqueda implica recuperar los documentos de un índice a través de un ‘IndexSearcher’.

- **Campos (Fields)**

Un documento consta de uno o más campos. Un campo es simplemente un par ‘nombre-valor’. Por ejemplo, un campo común que se encuentra en las aplicaciones es el título. En el caso de un campo de título, el nombre del campo es el título y el valor es el título de ese elemento de

contenido. La indexación en Lucene por lo tanto implica la creación de documentos que comprenden uno o más campos, y la adición de estos documentos a un 'IndexWriter'.

- **Búsqueda**

La búsqueda requiere un índice que ya esté construido. Se trata crear una consulta (por lo general a través de un 'QueryParser') y la entrega de esta consulta a un 'IndexSearcher[34]', el cual devuelve una lista de 'hits'.

- **Consultas**

Lucene tiene su propio mini-lenguaje para la realización de búsquedas. El lenguaje de consulta de Lucene permite al usuario especificar qué campo(s) va a buscar, los campos a los que se dará más peso, la posibilidad de realizar consultas booleanas (AND, OR, NOT) y otras funcionalidades.

8.1.1.2 Características

Lucene ofrece funciones a través de un API sencilla (Lucene, Apache Lucene - Features, 2010):

- Indexación escalable y de alto rendimiento
- Algoritmos de búsqueda
- Búsqueda de ranking: los mejores resultados son devueltos primero.
- Varios tipos de consulta: consultas por frase, por comodín (wilcard), consultas por proximidad, búsqueda por rangos, etc.
- Búsqueda por campos, (por ejemplo: autor, título, contenido).
- Búsqueda por rango de fechas.
- Ordenar por cualquier campo.
- Múltiples índices de búsqueda con resultados combinados.
- Solución multiplataforma

8.1.1.3 Comodines de Lucene

Los comodines (wildcards) que están disponibles en Lucene son (LuceneWiki, 2010):

- Consulta de prefijo: Por ejemplo 'libro *', que encontrará los documentos que contengan los términos tales como libro, librería, folleto, etc.
- Consulta Wild card: Permiten colocar un comodín en el centro del término de la consulta. Por ejemplo se pueden realizar búsquedas como: 'mi*pelling'. Esto coincidirá con 'misspelling', que es la forma correcta de escribir dicha palabra.

[34] Recupera los documentos de un índice

- Consulta de comodín (wildcard): Otro comodín que se puede usar es '?' (signo de interrogación); El '?' coincidirá con un solo carácter, esto permitirá realizar consultas como por ejemplo: Bra?il, La consulta coincidirá tanto con Brasil y Brasil.

8.1.2 Regain

Regain es un motor de búsqueda, similar a los motores de búsqueda web como Google, con la diferencia de que no busca en toda la web, sino que lo hace en sus propios archivos y documentos. El crawler de regain recupera sus archivos o páginas web, extrae todo el texto y lo coloca en un índice de búsqueda inteligente; todo este proceso es transparente para el usuario. Existen dos versiones de regain: ¡) La búsqueda de escritorio (The desktop search), la cual se va a utilizar en una computadora de escritorio normal, la misma que ofrece una búsqueda rápida de documentos o páginas web de la intranet. ii) El servidor de búsqueda (server search), esta se puede instalar en servidores web, proporciona funcionalidad para la búsqueda de un sitio web o una intranet servidora de archivos.

Regain está escrito en el lenguaje JAVA, y por lo tanto es aplicable a todas las plataformas compatibles con JAVA. El servidor de búsqueda funciona con JSP (Java Server Pages) y una biblioteca de etiquetas; la búsqueda de escritorio viene con su propio 'servidor web'. Regain está basado en LUCENE, una biblioteca para la creación y búsqueda de índices de búsqueda. Regain es liberado bajo licencia LGPL (Licencia pública general); es decir, regain puede ser utilizado de forma gratuita sin ningún límite temporal, se concede la autorización de usar, desarrollar y personalizar regain, con la única condición de que se proporcione el código fuente libremente para todo el mundo. (Schneider & Tesche, Regain, 2010)

	regain desktop search	regain server search
Best choice for newbies	✔	✘
regain Crawler	✔	✔
regain Search mask	✔	✔
Complete configuration of the crawler in one XML file	✔	✔
Simple configuration of the crawler over a web interface	✔	✘
Complete configuration of the search mask in one XML file	✔	✔
Customizable look of the search mask by JSPs	✔	✔
Integrated web server	✔	✘
Runs in a servlet engine (e.g. Tomcat)	✘	✔
Integration in the task bar	✔	✘

Figura 8.1 Comparación de las variantes de Regain (Schneider & Tesche, Regain manual, 2009).

8.1.2.1 Búsqueda con Regain

El trabajo de regain se divide en dos partes (Schneider & Tesche, Regain, 2010):

a) La creación del índice de búsqueda
b) La búsqueda en el índice de búsqueda.

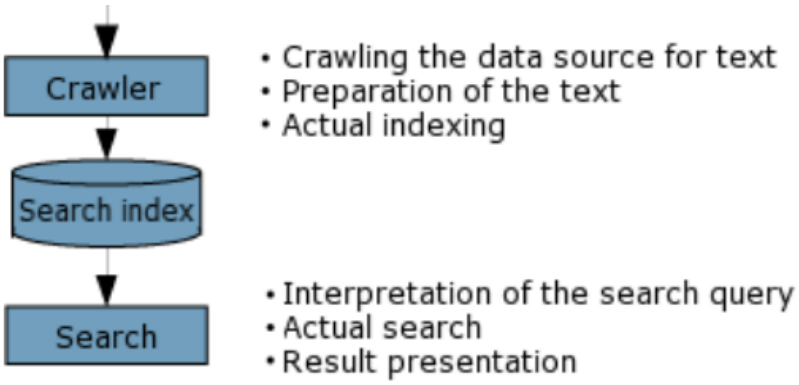

Figura 5.3. Cómo realiza las búsquedas regain (Schneider & Tesche, Regain, 2010).

a) La creación del índice de búsqueda

El crawler (rastreador) realiza la búsqueda en un sitio web o en un árbol de directorio de documentos. En la configuración de regain se puede especificar exactamente lo que debe ser rastreado. El texto de cada documento se extrae mediante los denominados 'preparadores' y el texto se agrega al índice de búsqueda. Los preparadores son los responsables del formato de cada documento. Por ejemplo, hay un preparador HTML que es capaz de leer el texto de documentos HTML, un preparador PDF que es capaz de leer archivos PDF y así sucesivamente.

Lista de preparadores disponibles en Regain (Schneider & Tesche, Regain, 2010):

- HtmlPreparator.- preparador para los archivos HTML.
- PdfBoxPreparator.-para archivos de documentos portátiles (PDF), basado en PDFBox.
- OpenOfficePreparator.- para los archivos de OpenOffice.
- PoiMsOfficePreparators.- un conjunto de preparadores escrito en Java, utilizando POI[35] para archivos de MS Office (Word, Excel y Powerpoint).
- PlainTextPreparator.- para archivos simples de texto ASCII.
- XmlPreparator.- para los archivos XML.
- SimpleRtfPreparator - preparador para formato de archivos de texto rico (RTF), propio de Regain.
- SwingRtfPreparator.- otro preparador-RTF basado en JEditorPane del marco para Java Swing.

[35] API de java para documentos de Microsoft.

- EmptyPreparator.-para almacenar url y la fecha solamente, ignora el contenido del archivo.
- ExternalPreparator.- un preparador genérico, que llama a cualquier programa o script externo para la preparación.

También existen dos preparadores más que están disponibles solo para sistemas MS Windows:

- IfilterPreparator.- un preparador en la parte superior de la MS IFilter-API. Una gama de formatos de archivo apoyados que dependen de los IFilter que están instalados en el sistema. Sin embargo, el propio Windows ya tiene un número de IFilters.
- JacobMsOfficePreparators.- este es un conjunto de preparadores para Word, Excel y PowerPoint los cuales proporcionan un buen uso de la extracción de contenido. Sin embargo, una instalación de MS Office es necesaria, porque los programas de Office están involucrados.

b) Búsqueda en el índice de búsqueda

Después de haber creado un índice de búsqueda, regain está en condiciones de realizar las búsquedas. El índice de búsqueda almacena los datos acerca de los documentos de tal forma que se podrán encontrar documentos que contengan una determinada palabra clave rápidamente en las solicitudes de búsqueda. Regain utiliza Lucene para la creación del índice y la búsqueda basada en el índice. Lucene separa los datos de un documento en varios campos clasificados; gracias a esto también se pueden especificar que campos se va a consultar. Por ejemplo realizar búsquedas en los documentos PDF: “extension:pdf”.

8.1.2.2 Campos Estándar de Regain

Regain crea los siguientes campos

- Url.- La URL del documento.
- Content.- El texto del documento extraído por los preparadores.
- Title.- El título del documento (si tiene).
- Summary.- El resumen; se muestra en la lista de resultados.
- Headlines.- Los titulares (si los hay) que figuran en el documento.
- Size.- El tamaño del documento en bytes (no pueden ser consultados).
- Last-modified.- La fecha del último cambio en el formato AAAA-MM-DD HH:MM (no pueden ser consultados).
- Path.- La ruta de navegación para el documento. (No pueden ser consultada)
- Groups.- Contiene los grupos de usuarios que tienen permiso para leer el documento. Sólo se establece cuando la gestión de los derechos de acceso está habilitada.
- Otros campos pueden añadirse.- La configuración por defecto añade una extensión de campo de almacenamiento de archivo del documento (por ejemplo, pdf).

8.1.2.3 Otras Características de Regain

A continuación se mencionan otras de las características de Regain (Schneider & Tesche, Regain, 2010):

Búsqueda

Regain utiliza la sintaxis de búsqueda de Lucene. Por lo tanto, es posible expresar consultas de búsqueda muy específicas. Las posibilidades más importantes son las siguientes:

- Los operadores booleanos
- Wildcards (Los comodines)
- Phonetic search (Búsqueda fonética)
- Grouping (Agrupamiento)

Definir el espacio de búsqueda

- Usando Regain se puede especificar con mucha exactitud lo que debe ser indexado y lo que no.
- Lista blanca y negra: Con las denominadas listas blancas y negras, se puede aislar exactamente qué documentos debe procesar el crawler. Por ejemplo, se podrán indexar todos los documentos de'http://www.murfman.de' a excepción de los documentos en 'http://www.murfman.de/dynamiccontent'.
- Varias fuentes en una lista: Se pueden indexar documentos de diferentes sistemas de archivos y/o sitio web en un mismo índice de búsqueda.
- Indexación parcial: Se supone que el índice de búsqueda contiene los elementos de una unidad de red (servidor de archivos) y una página web. Regain puede actualizar solo los documentos de la unidad de red; de esta manera se pueden actualizar solo algunas unidades cada hora y otras solo cada semana.

Indexación

- Despliegue en caliente: Cambio de un nuevo índice de búsqueda sin necesidad de reiniciar el motor de servlets (por ejemplo tomcat).
- Lista de palabras de parada (stopword list): Definir las palabras que no deben ser indexadas.
- Análisis de archivos: Si se desea, todos los pasos intermedios del proceso de indexación pueden ser escritos como archivos. De este modo se puede ver exactamente lo que se interpone en el índice de búsqueda.
- Extracción de contenido para HTML: El índice solo recupera el contenido real de las páginas web, se elimina el pie de página para el archivo.
- Puntos de interrupción: El crawler crea periódicamente puntos de interrupción, al hacerlo, el estado actual del índice de búsqueda se copia en un directorio independiente. Si la actualización del índice es cancelada (por ejemplo si el computador se apaga), el crawler volverá al último punto de interrupción creado.

Rating de los resultados de búsqueda

Los resultados de la búsqueda se clasifican de acuerdo a la frecuencia relativa con que los términos aparecen en un documento. Por ejemplo, un documento con 100 palabras que contiene el término de búsqueda 5 veces será clasificado como un éxito frente a un documento de 1000 palabras que contiene el término de búsqueda 10 veces.

8.1.3 Nutch

Nutch es un software de búsqueda-web de código abierto; es un esfuerzo por construir un motor de búsqueda libre y de código abierto. Utiliza Lucene para buscar e indexar componentes. El robot de búsqueda ha sido escrito desde cero exclusivamente para este proyecto. Está basado en LUCENE y en SOLR, pero agregando características web específicas, tales como un crawler (rastreador), una base de datos de enlace-gráfico, analizadores de HTML y otros formatos de documentos, etc. Nutch puede ejecutarse en una máquina simple, pero se puede explotar mucho más su potencial si se ejecuta en un cluster Hadoop[36]. El sistema de nutch puede ser mejorado utilizando un mecanismo de plugin (Nutch, 2010) (NutchWiki, 2010).

Características: (NutchWiki, 2010)

- Obtención, análisis e indexación en paralelo y/o distribuido.
- Plugins.
- Muchos Formatos: texto plano, HTML, XML, ZIP, OpenDocument, Microsoft Office (Word, Excel, Powerpoint), PDF, Javascript , RSS, RTF, MP3 (etiquetas ID3)
- Ontología
- Clustering
- MapReduce [37]
- Sistema de fichero distribuido (a través de Hadoop)
- Base de datos de enlace-gráfico.
- Autenticación NTLM

8.1.3.1 Requerimientos para la Instalación

Los requerimientos tanto de software como de hardware se mencionan a continuación (NutchWiki, 2010):

[36] El proyecto Apache Hadoop desarrolla software de código abierto para la informática fiable, escalable y distribuida. http://hadoop.apache.org/

[37] Permite a conjuntos de datos masivos ser procesados de forma distribuida, rompiendo la transformación en muchos cálculos pequeños.http://wiki.apache.org/nutch/MapReduce

Software necesario

Ya que Nutch está escrito en Java, es posible tener trabajando a Nutch en Windows siempre y cuando el software necesario este instalado. A continuación se enumera el software que se necesitará para poder utilizar Nutch:

1. Java (JRE)
2. Cygwin
3. Tomcat
4. Apache Nutch

El proceso de instalación se estos componentes se explica en el anexo 2.

Requerimientos de Hardware

En realidad, el acarreo y las actualizaciones de las bases de datos requieren un gran número de discos, y la búsqueda será más rápida con más memoria RAM. Pero esto depende de que tan grande se esté intentando construir el índice.

Requisitos para la indización

Como regla general, cada página a buscar requiere alrededor de 10k de disco (para el caché de la página, el texto, el índice, las entradas de DB, etc.). Así, para un millón de páginas se requerirá de un terabyte (1 Tb) de almacenamiento.

Requisitos para la búsqueda

Si el tráfico de punta es alrededor de solo una consulta por segundo, entonces se podrán tener 20 millones o más de páginas por nodo. Para un tráfico pico más alto, se debería aumentar la memoria RAM

8.1.3.2 Componentes Principales de Nutch

El motor de búsqueda Nutch consiste a grandes rasgos de tres componentes (NutchWiki, 2010):

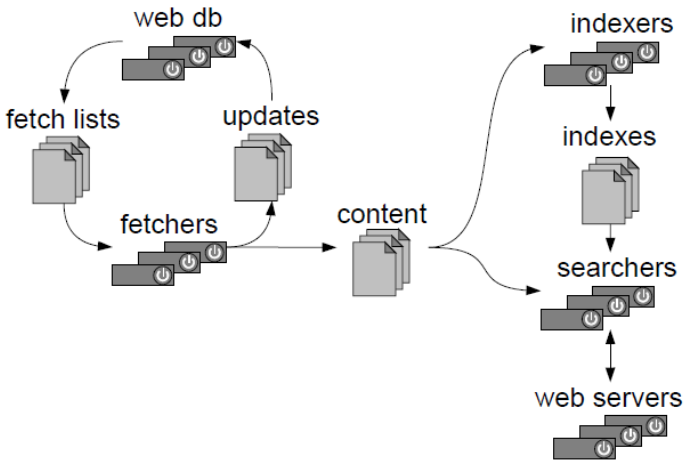

Figura 8.2 Arquitectura de Nutch (Cutting, 2004).

1. El crawler, que descubre y recupera las páginas web.
2. El 'WebDB', una base de datos personalizada que almacena las direcciones URL conocidas y obtiene el contenido de las páginas.
3. El 'Indexer', que analiza las páginas y genera índices basadas en las palabras clave de las mismas.

8.1.3.3 La Estructura del Índice

La estructura del índice que se forma después de la indexación es la siguiente (NutchWiki, 2010):

- Boost
- digest
- lang
- segment
- tstamp
- anchor
- title
- site
- host
- url
- content
- lastModified
- date
- contentLength
- type
- primaryType
- subType
- domain
- tld
- category
- subcollection
- site
- host
- url
- content

8.1.3.4 Plugins Disponibles con Nutch

Nutch dispone de los siguientes plugins para analizar los documentos (NutchWiki, 2010):

- **clustering-carrot2:** Resultados de la búsqueda en línea usando clústeres de componentes Carrot2.
- **creativecommons:** Soporte para el rastreo y búsqueda de contenido licenciado bajo Creative-commons.
- **index-basic:** Agrega el URL, el contenido y los campos al índice.
- **index-more:** Agrega la fecha, la longitud del contenido, y campos como contentType, primaryType y campos de subtipo al índice.
- **languageidentifier:** Agrega el campo 'lang' al índice y permite realizar consultas sobre este.
- **Ontology:** Ayuda a refinar consultas basadas en los archivos owl.
- **parse-ext:** Un contenedor que invoca un comando externo para hacer el trabajo de análisis real.
- **parse-html:** Analiza los documentos HTML
- **parse-js:** Analiza Java Script
- **parse-mp3:** Analiza mp3
- **parse-zip:** Analiza archivos ZIP
- **parse-mspowerpoint:** Analiza archivos de Microsoft PowerPoint
- **parse-msword:** Analiza documentos de MS Word

- **parse-msexcel:** Analiza documentos de MS Excel
- **parse-pdf:** Analiza archivos PDF
- **parse-rss:** Analiza los feeds RSS
- **parse-oo:** Analiza archivos OpenOffice
- **parse-swf:** Analiza Shockwave Flash
- **parse-rtf:** Analiza archivos RTF
- **parse-text:** Analiza documentos de texto
- **protocol-file:** Recupera documentos desde el sistema de archivos
- **protocol-ftp:** Recupera documentos a través de ftp
- **protocol-http:** Recupera documentos a través de HTTP
- **protocol-httpclient:** Recupera documentos a través de HTTP y HTTPS
- **query-basic:** Ejecuta consultas sobre los campos 'content', 'url' y 'anchor fields'
- **query-more:** Ejecuta las consultas para los siguientes campos: date, content-length, contentType, primaryType and subType fields.
- **query-site:** Ejecuta consultas en el campo 'site'.
- **query-url:** Ejecuta consultas en el campo 'url'.

8.2 Anexo 2 - Instalación de Nutch

Se tomo como base para la siguiente instalación: (Wang, 2008), (Garcia, 2010).

8.2.1 Software Necesario

Para seguir los pasos descritos en el artículo se necesitará (se indican las versiones usadas durante la elaboración del artículo, pero no tienen por qué ser exactamente las mismas):

- Nutch 1.2
- Java JDK 1.6 .0_21
- Apache Tomcat 6.0.29
- Cygwin, sólo si estás en Windows (con Linux no se necesita)

Nutch 1.2

El software se puede descargar desde la página http://nutch.apache.org/ , la versión utilizada en este estudio de tesis es la versión 1.2 (Octubre 2010)

Java JDK

La versión del JDK sobre la cual se trabajo es la 1.6.0_21

Apache Tomcat

El software se puede descargar desde la página http://tomcat.apache.org/ , se pueden encontrar varias versiones de Tomcat, siendo la más reciente la versión 7.0.0 (Agosto 2010); pero ésta aún se encuentra en versión beta, por lo que se descargó la última versión estable correspondiente a la versión 6.0.29

Tomcat 6.0.x

Spec versions:	Servlet 2.5, JSP 2.1
Stable:	Yes
Enhancements:	Yes
Bug Fixes:	Yes
Security Fixes:	Yes
Releases:	Yes
Release Manager:	Jean-Frederic Clere (jfclere)
Process:	RTC
Listed on download pages:	Yes

Tomcat 7.0.x

Spec versions:	Servlet 3.0, JSP 2.2, EL 2.2
Stable:	No
Enhancements:	Yes
Bug Fixes:	Yes
Security Fixes:	Yes
Releases:	Yes
Release Manager:	Mark Thomas (markt)
Process:	CTR
Listed on download pages:	Yes

Figura 8.3 Comparación entre las versiones de Tomcat.

Cygwin

El software se puede descargar desde la página http, la última versión hasta la fecha (Agosto 2010) fue la instalada para este estudio, correspondiente a la versión 1.7.5

Cygwin es como un entorno de Linux para Windows. Se compone de dos partes:

- Una DLL (cygwin1.dll), que actúa como una capa de emulación Linux API proporcionando importantes funcionalidades de la API de Linux.
- Una colección de herramientas que proporcionan apariencia Linux.

8.2.2 Configuración de Nutch 1.2

La instalación del JDK, del Tomcat y de Cygwin, deberían instalarse sin ningún problema, tal solo siguiendo los pasos de la instalación; una vez instalados estos tres componentes, realizaremos la instalación del nutch, para esto seguimos los siguientes pasos:

1. Descomprimimos el fichero 'nutch-1-0.tar.gz' en un directorio cualquiera de nuestro computador, en este caso fue en 'C:\nutch-1.2'.
2. Una vez descomprimido se debe configurar el crawler, para esto ejecutamos el programa cygwin y mediante la consola que aparece nos ubicamos en el directorio donde descomprimimos en nutch.

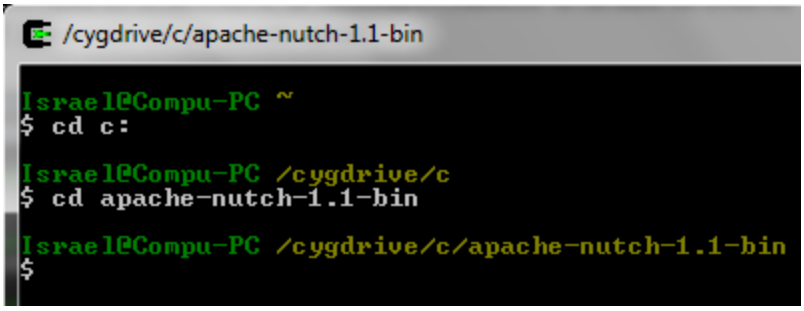

Figura 8.4 Configuración de Nutch mediante cygwin.

3. Una vez posicionados en el directorio, se necesitará establecer la variable de entorno 'JAVA_HOME'. (La versión del jdk puede variar)

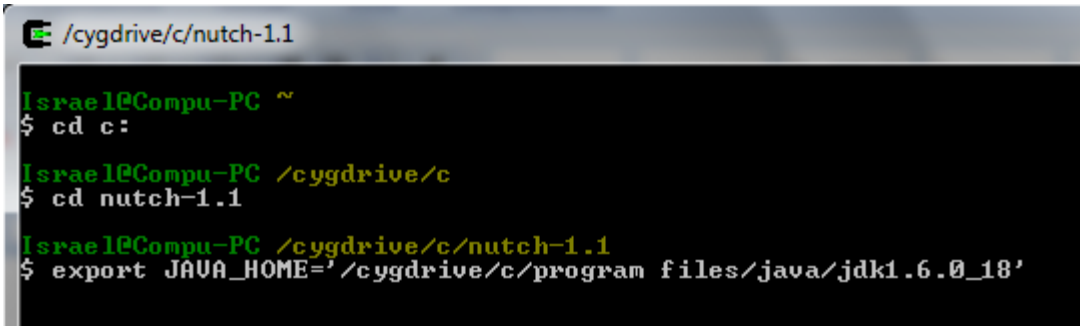

Figura 8.5 Configuración de Nutch mediante cygwin - Variable de entorno.

Una alternativa para establecer esta variable de entorno permanentemente en un sistema Windows, es creando una nueva variable de entorno llamada 'JAVA_HOME' cuyo valor será la ruta del directorio donde se encuentre instalado el jdk.

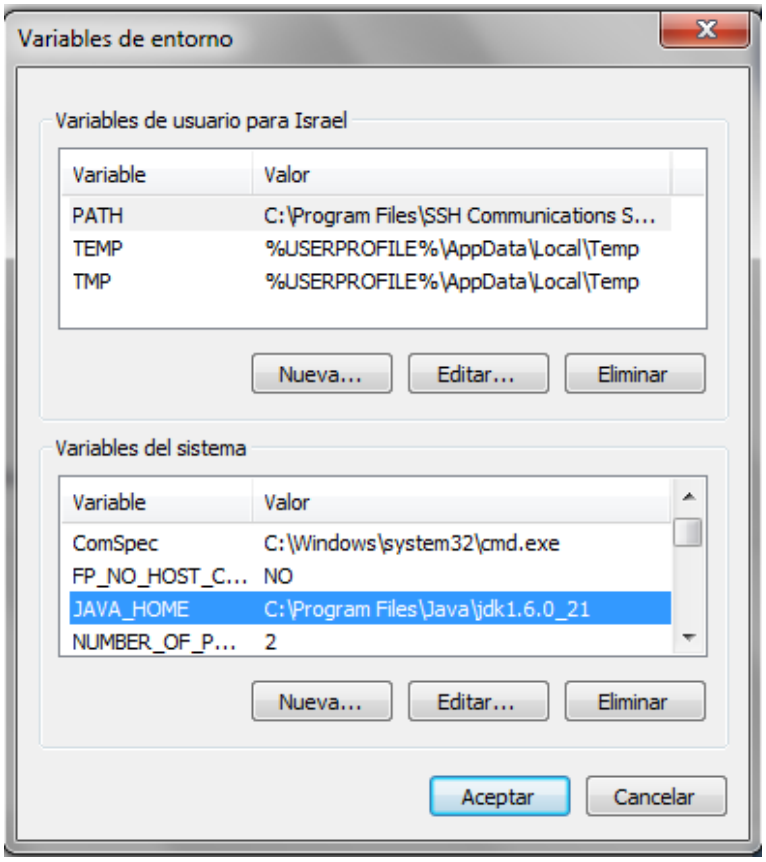

Figura 8.6 Establecimiento de variable de entorno desde Windows.

4. Crear una nueva carpeta llamada 'urls' (puede ser cualquier nombre) dentro del directorio donde se descomprimió el nutch (En este caso en 'C:\nutch-1.2'); una vez realizado esto, dentro del directorio llamado 'urls' crear un fichero .txt (en este caso llamado '*otech_crawl.txt*'); y dentro del fichero .txt colocar las páginas en las cuales va a trabajar el crawler. Si queremos colocar más de una dirección, tendremos que poner una por línea.

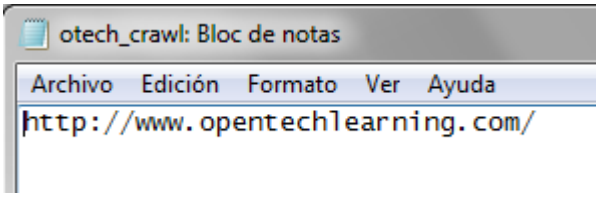

Figura 8.7 Establecimiento de las páginas a indexar.

5. Para comprobar que la carpeta urls se copio correctamente, en la línea de comandos de cygwin, escribimos el comando 'ls', y verificamos que exista la carpeta que creamos llamada urls.

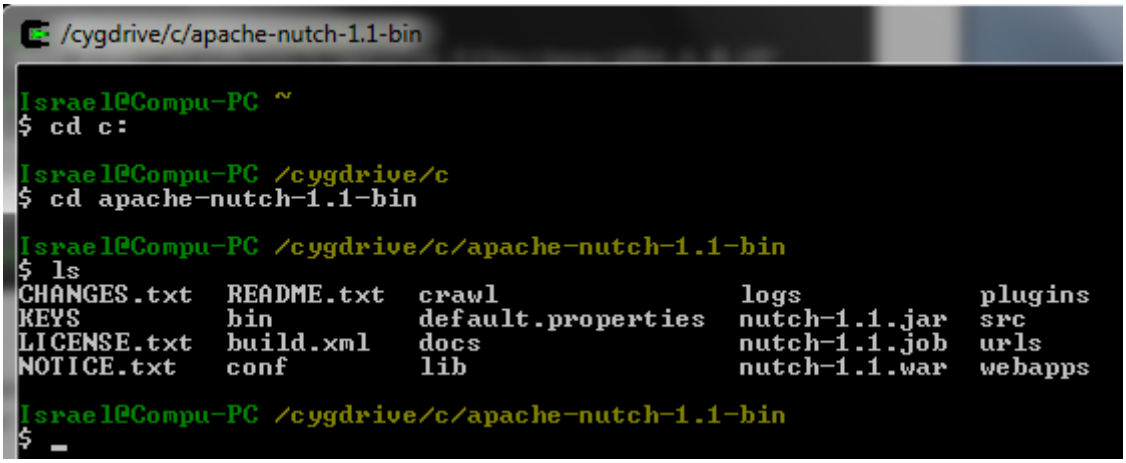

Figura 8.8 Configuración de Nutch mediante cygwin (2).

6. Dentro del directorio donde se instalo el nutch ('C:\nutch-1.2'),localizamos en la carpeta 'conf' los siguientes ficheros, que son los que usaremos para configurar el crawler:
   - **crawl-urlfilter.txt:** este fichero contiene expresiones regulares prefijadas con los caracteres '+' y '-' para indicar si las urls que cumplan el patrón indicado son incluidas o ignoradas.
   - **nutch-site.xml:** en este fichero indicaremos algunas propiedades de carácter general.

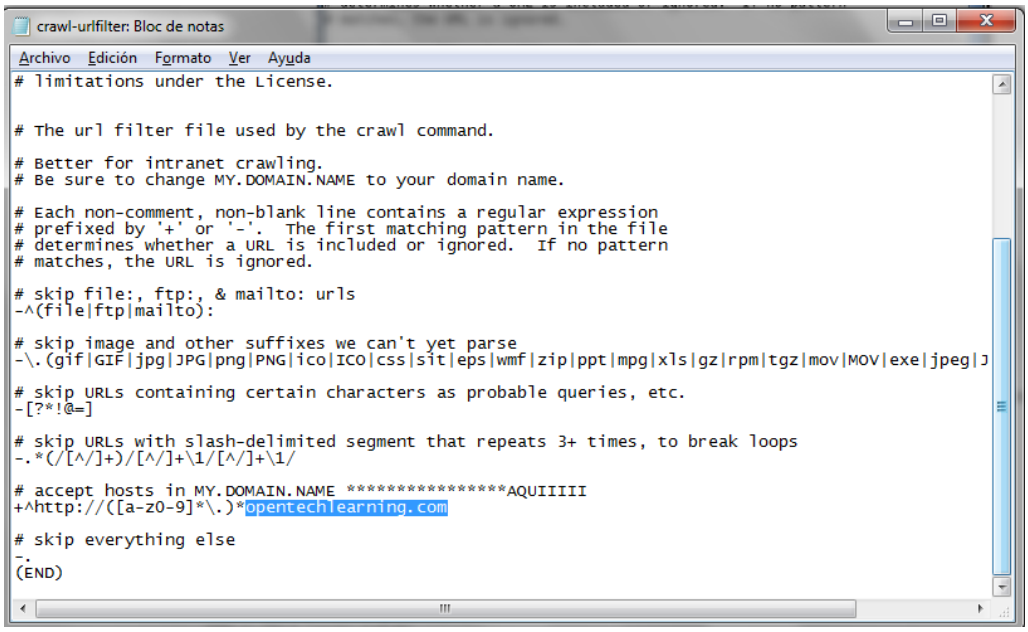

Figura 8.9 Configuración del archivo crawl-urlfilter.txt

***crawl-urlfilter.txt**. De momento, con cambiar 'MY.DOMAIN.NAME' por el dominio sobre el cual va a trabajar el crawler basta. En este ejemplo, se indexara el dominio opentechlearning.com.*

7. En el fichero **nutch-site.xml**, se deberá definir al menos estas tres propiedades:

```
<property>
<name>http.agent.name</name>
<value>OTech</value>
<description>Open Tech </description>
</property>
<property>
<name>http.agent.description</name>
<value>Otech crawler</value>
<description>Open Tech Crawler</description>
</property>
<property>
<name>http.agent.url</name>
<value>http://www.opentechlearning.com</value>
<description>http://www.opentechlearning.com</description>
</property>
<property>
<name>http.agent.email</name>
<value> admin@opentechlearning.com </value>
<description> admin@opentechlearning.com </description>
</property>
```

Figura 8.10 Configuración del archivo nutch-site.xml

8. Una vez realizados estos pasos, estamos en condiciones de ejecutar el crawler. Desde el Cygwin, ejecutamos desde el directorio donde se descomprimió el Nutch, lo siguiente:

bin/nutch crawl urls -threads 5 -dir crawl -depth 3 -topN 10

Con estos parámetros se va a lanzar el crawler sobre las urls indicadas en el fichero de texto dentro del directorio urls, con 5 hilos de ejecución (-threads), generará el índice en el directorio crawl (-dir), con una profundidad desde la página raíz de 3 niveles (-depth) y un máximo de 10 páginas por nivel (-topN).

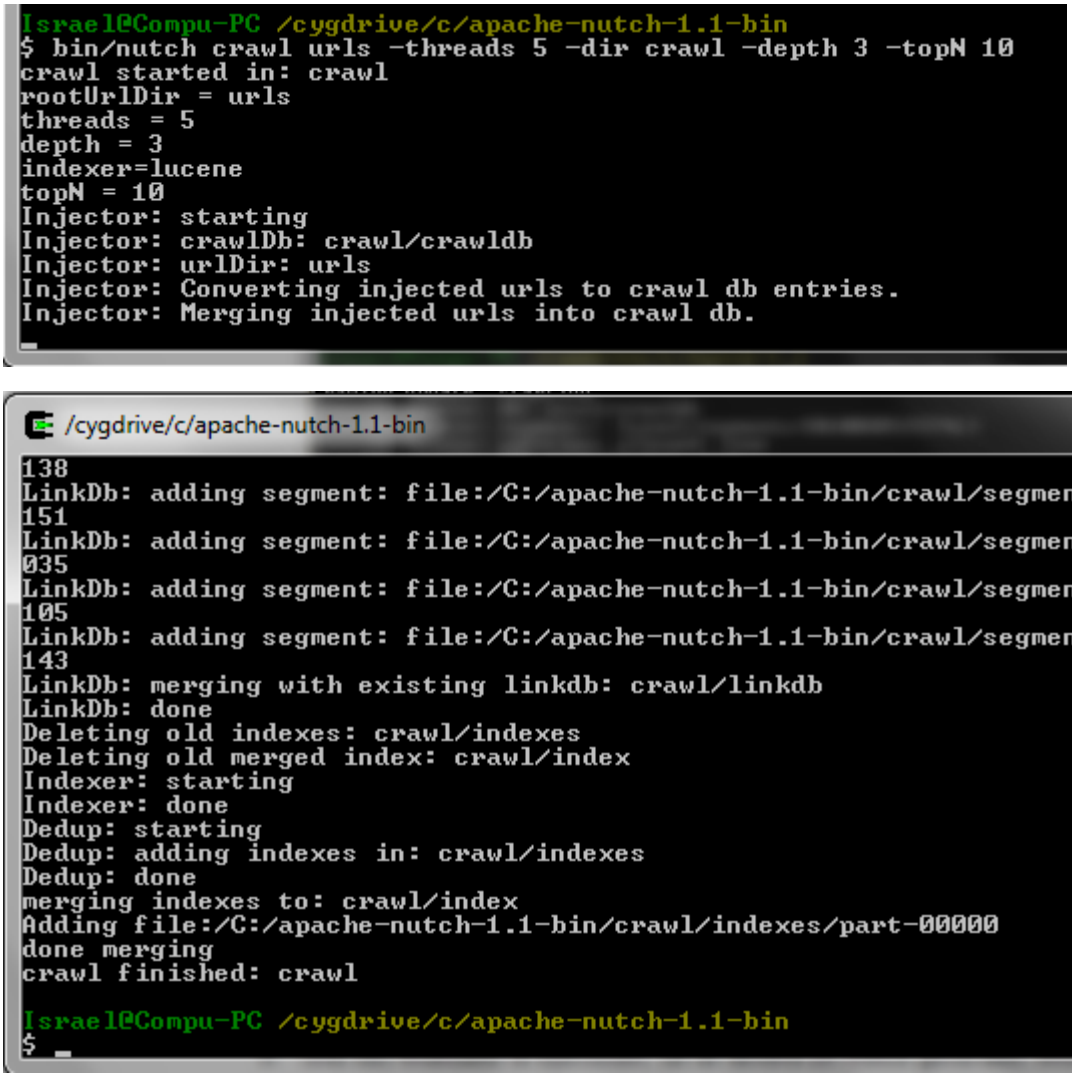

Figura 8.11 Ejecución del crawl desde cygwin.

9. Una vez finalizada la ejecución, ya se tendrá un índice generado. Ahora se necesita realizar la búsqueda; aquí es donde entra en juego el Tomcat que hemos instalado.
   En el directorio donde se encuentra el Nutch, se encuentra el fichero nutch-1.2.war, que será el que desplegaremos en el Tomcat y que nos permitirá realizar búsquedas sobre el índice. Para desplegarlo, simplemente copiar este fichero en el directorio webapps del Tomcat (reiniciar el Tomcat, si está iniciado).

10. Modificar el fichero 'nutch-site.xml', en el directorio webapps/nutch-1.2/WEB-INF/classes/ del Tomcat para indicarle a la aplicación web dónde está el índice (será el directorio indicado en el parámetro -dir cuando ejecutamos el crawler). En nuestro caso, se incluyó en el fichero nutch-site.xml lo siguiente:

```
<configuration>

<property>
<name>searcher.dir</name>
<value>C:\apache-nutch-1.1-bin\crawl</value>
</property>

</configuration>
```

Figura 8.12 Configuración del archivo 'nutch-site.xml'.

11. Una vez realizado todo esto, podemos realizar las búsquedas mediante nutch, escribimos en el navegador: 'http://localhost:8080/nutch-1.2/''

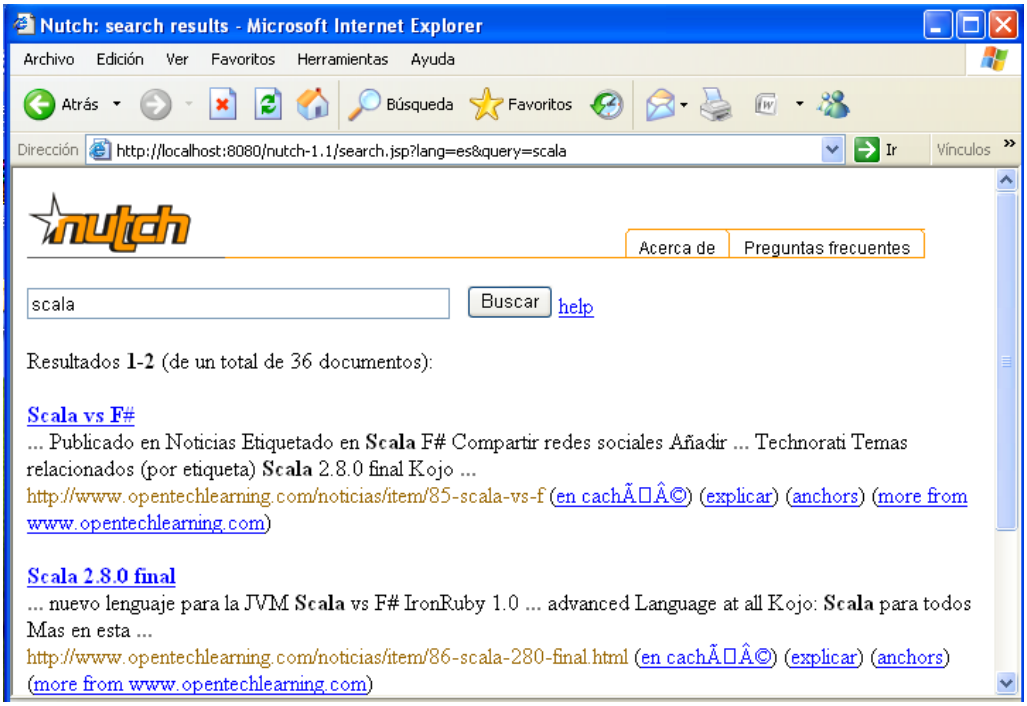

Figura 8.13 Interfaz de consulta de Nutch.

8.3 Anexo 3 - Instalación Regain

Existen dos versiones de Regain, 1) servidor de búsqueda, y 2) búsqueda de escritorio; esta última es una versión muy fácil de instalar para entornos Windows gracias a su asistente de instalación. La versión instalada en este estudio de tesis es la búsqueda de escritorio.

Regain es de código abierto, y se puede encontrar sus diferentes versiones en: http://regain.sourceforge.net/download.php. La versión instalada es la última versión hasta la fecha (Agosto 2010) y corresponde a la versión 1.6.6.

1. Descargamos 'Regain Desktop Search 1.6.6 for windows' (Installer, version .exe).
2. Ejecutamos el asistente de instalación:

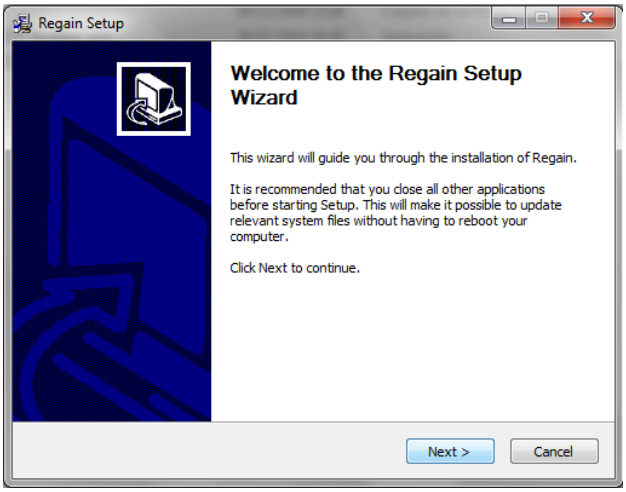

Figura 8.14 Asistente instalación de Nutch (1).

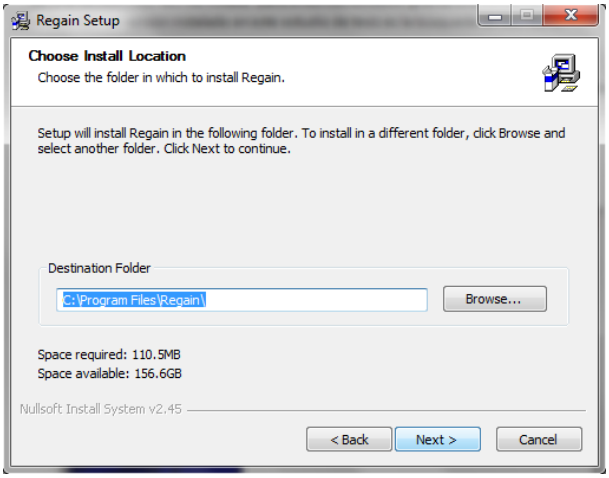

Figura 8.15 Asistente instalación de Nutch (2).

3. Al momento de ejecutar 'Regain.exe', aparece en la barra de herramientas el ícono de regain, esto nos indica que la instalación ha sido un éxito.

Figura 8.16 Símbolo de Regain en la barra de herramientas de Windows.

4. La configuración de regain es muy sencilla, en preferencias podemos elegir que directorios locales se van a indexar, que páginas web se van a indexar; así mismo podemos excluir dichos directorios y páginas.

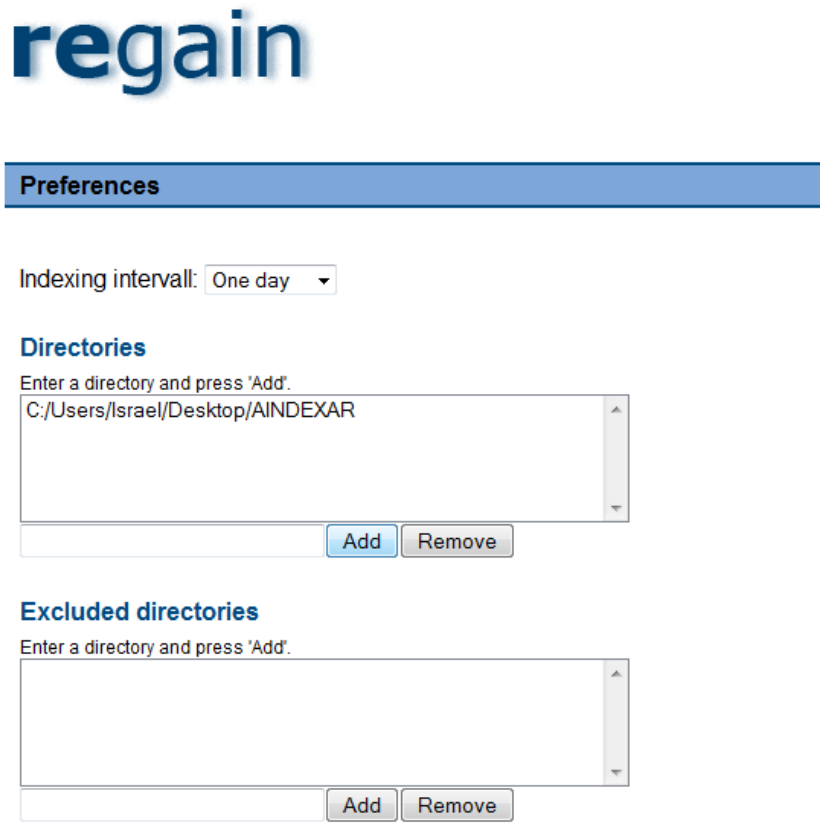

Figura 8.17 Configuración de Regain - Directorios a indexar.

Figura 8.18 Configuración de Regain - Páginas a indexar.

8.4 Anexo 4 - Disponibilidad de Cursos en los Repositorios Seleccionados

Tabla 8-1 Categorías de recursos disponibles en repositorio EduTube.

EduTube		
Categoría	**Subdivisión**	
Arts & Humanities	Art	
	Film & Media	
	History	
	Language	
	Music	
Sciences	Biology	
	Chemistry	
	Mathematics	
	Physics	
	Space & Astronomy	
Applied Sciences	Agriculture	
	Computer Science	
	Education	
	Software	
	Technology	
Social Sciences	Anthropology	
	Development	
	Environment	
	Politics	
	Psychology	
Other	Animals & Wildlife	Health
	Archaeology	Law
	Architecture	Literature
	DIY & How-To	Marketing
	Economics	Math & Statistics
	Engineering	Philosophy
	Food & Cooking	Religion
	Geography	Sociology
	Geology & Earth	Sports & Martial Arts

Tabla 8-2 Categorías de recursos disponibles en repositorio MIT.

OPENCOURSEWARE CONSORTIUM (MIT)	
CURSOS	**DEPARTAMENTO**
Architecture and Planning	Architecture Media Arts and Sciences Urban Studies and Planning
Engineering	Aeronautics and Astronautics Biological Engineering Chemical Engineering Civil and Environmental Engineering Electrical Engineering and Computer Science Engineering Systems Division Materials Science and Engineering Mechanical Engineering Nuclear Science and Engineering
Health Sciences and Technology	Health Sciences and Technology
Humanities, Arts, and Social Sciences	Anthropology Comparative Media Studies Economics Foreign Languages and Literatures History Linguistics and Philosophy Literature Music and Theater Arts Political Science Science, Technology, and Society Women's and Gender Studies Writing and Humanistic Studies
Management	Sloan School of Management
Science	Biology Brain and Cognitive Sciences Chemistry Earth, Atmospheric, and Planetary Sciences Mathematics Physics

Other Programs	Athletics, Physical Education and Recreation Special Programs
Supplemental Resources	Supplemental Resources
Cross-Disciplinary Topics	Energy Courses Environment Courses

Tabla 8-3 Categorías de recursos disponibles en repositorio Merlot

MERLOT		
Categoría	**Subdivisión**	
Academic Support Services	Accessibility and Assistive Technology	LMS Planning and Implementation
	Accreditation	Library and Information Services
	Content Repositories	Mobile Computing
	Course Redesign	Online Degree Program Development
	Degree Completion/Accelerating Graduation	Scholarship of Teaching and Learning
	EPortfolios	Smart Classrooms
	Faculty Development	Staff Development
	Hybrid and Online Course Development	Textbook Affordability
	Institutional Research on Technology Use	Virtual Environments
	K-20 Initiatives	other Innovations/Emerging Technologies
Arts	Art History	General
	Cinema	Music
	Dance	Photography
	Fine Arts	Theatre
Business	Accounting	General
	Business Law	Information Systems
	Corporate Social Responsibility	International Business
	E-Commerce	Management
	Economics	Marketing
	Finance	Professional Coaching
Education	Educational Leadership	
	General	
	TeacherEd	
Humanities	English	Philosophy

	General	Religion
	History	World Languages
Mathematics and Statistics	Mathematics	
	Statistics and Probability	
Science and Technology	Agriculture	General Science
	Astronomy	Geoscience
	Biology	Health Sciences
	Chemistry	Information Technology
	Communication Sciences and Disorders	Nanotechnology
	Computer Science	Physics
	Engineering	Technical Allied Health
	Fire Safety	
Social Sciences	Anthropology	Psychology
	Criminal Justice	Sociology
	General	Sports and Games
	Geography	Statistics
	Law	Womens Studies
	Political Science	

Tabla 8-4 Categorías de recursos disponibles en repositorio Connexions.

CONNEXIONS	
Categorías	**Observación**
Arts	Modules: 1216, Collections: 69
Business	Modules: 544, Collections: 47
Humanities	Modules: 1841, Collections: 150
Mathematics and Statistics	Modules: 3709, Collections: 111
Science and Technology	Modules: 5847, Collections: 392
Social Sciences	Modules: 1901, Collections: 135

8.5 Anexo 5 - Archivos Disponibles para la Categoría "Anthropology" en cada Repositorio.

Tabla 8-5 Categoría de antropología en repositorio MIT.

MIT		
ID	**Título**	**Descripción**
MIT001	American Dream: Exploring Class in the U.S.	Americans have historically preferred to think of the United States in classless terms, as a land of economic opportunity equally open to all. Yet, social class remains a central fault line in the U.S. Subject explores the experiences and understandings of class among Americans positioned at different points along the U.S. social spectrum. Considers a variety of classic frameworks for analyzing social class and uses memoirs, novels and ethnographies to gain a sense of how class is experienced in daily life and how it intersects with other forms of social difference such as race and gender.
MIT002	Anthropological Theory	This course introduces students to some of the major social theories and debates that inspire and inform anthropological analysis. Over the course of the semester, we will investigate a range of theoretical propositions concerning such topics as agency, structure, subjectivity, history, social change, power, culture, and the politics of representation. Ultimately, all theories can be read as statements about human beings and the worlds they create and inhabit. We will approach each theoretical perspective or proposition on three levels: (1) in terms of its analytical or explanatory power for understanding human behavior and the social world; (2) in the context of the social and historical circumstances in which they were produced; and (3) as contributions to ongoing dialogues and debate.
MIT003	Anthropology of the Middle East	This course examines traditional performances of the Arabic-speaking populations of the Middle East and North Africa. Starting with the history of the ways in which the West has discovered, translated and written about the Orient, we will consider how power and politics play roles in the production of culture, narrative and performance. This approach assumes that performance, verbal art, and oral literature lend themselves to spontaneous adaptation and to oblique expression of ideas and opinions whose utterance would otherwise be censorable or disruptive. In particular we will be concerned with the way traditional performance practices are affected by and respond to the consequences of modernization.Topics include oral epic performance, sacred narrative, Koranic chant performance, the folktale, solo performance, cultural production and resistance.
MIT004	Anthropology of War and Peace	This class has been reorganized to focus primarily on the War in Iraq. As in previous years, the class still examines war in cross-cultural perspective, asking whether war is intrinsic to human nature, what causes war, how particular cultural experiences of war differ, and how war has affected American culture.

MIT

ID	Título	Descripción
MIT005	Culture, Embodiment and the Senses	Culture, Embodiment, and the Senses will provide an historical and cross-cultural analysis of the politics of sensory experience. The subject will address western philosophical debates about mind, brain, emotion, and the body and the historical value placed upon sight, reason, and rationality, versus smell, taste, and touch as acceptable modes of knowing and knowledge production. We will assess cultural traditions that challenge scientific interpretations of experience arising from western philosophical and physiological models. The class will examine how sensory experience lies beyond the realm of individual physiological or psychological responses and occurs within a culturally elaborated field of social relations. Finally, we will debate how discourse about the senses is a product of particular modes of knowledge production that are themselves contested fields of power relations.
MIT006	Dilemmas in Bio-Medical Ethics: Playing God or Doing Good?	This course is an introduction to the cross-cultural study of bio-medical ethics. It examines moral foundations of the science and practice of western bio-medicine through case studies of abortion, contraception, cloning, organ transplantation, and other issues. It also evaluates challenges that new medical technologies pose to the practice and availability of medical services around the globe, and to cross-cultural ideas of kinship and personhood. It discusses critiques of the bio-medical tradition from anthropological, feminist, legal, religious, and cross-cultural theorists.
MIT007	Documenting Culture	How — and why — do people seek to capture everyday life on film? What can we learn from such films? This course challenges distinctions commonly made between documentary and ethnographic films to consider how human cultural life is portrayed in both. It considers the interests, which motivate such filmmakers ranging from curiosity about "exotic" people to a concern with capturing "real life" to a desire for advocacy. Students will view documentaries about people both in the U.S. and abroad and will consider such issues as the relationship between film images and "reality," the tensions between art and observation, and the ethical relationship between filmmakers and those they film.
MIT008	Environmental Struggles	This class explores the interrelationship between humans and natural environments. It does so by focusing on conflict over access to and use of the environment as well as ideas about "nature" in various parts of the world.
MIT009	Ethnic and National Identity	An introduction to the cross-cultural study of ethnic and national identity. We examine the concept of social identity, and consider the ways in which gendered, linguistic, religious, and ethno-racial identity components interact. We explore the history of nationalism, including the emergence of the idea of the nation-state, as well as ethnic conflict, globalization, identity politics, and human rights.

MIT

ID	Título	Descripción
MIT010	Gender, Power, and International Development	After decades of efforts to promote development, why is there so much poverty in the world? What are some of the root causes of inequality world-wide and why do poverty, economic transformations and development policies often have different consequences for women and men? This course explores these issues while also examining the history of development itself, its underlying assumptions, and its range of supporters and critics. It considers the various meanings given to development by women and men, primarily as residents of particular regions, but also as aid workers, policy makers and government officials. In considering how development projects and policies are experienced in daily life in urban and rural areas in Africa, Latin America, Asia and Melanesia, this course asks what are the underlying political, economic, social, and gender dynamics that make "development" an ongoing problem world-wide.
MIT011	Gender, Sexuality, and Society	This course seeks to examine how people experience gender - what it means to be a man or a woman - and sexuality in a variety of historical and cultural contexts. We will explore how gender and sexuality relate to other categories of social identity and difference, such as race and ethnicity, economic and social standing, urban or rural life, etc. One goal of the class is to learn how to critically assess media and other popular representations of gender roles and stereotypes. Another is to gain a greater sense of the diversity of human social practices and beliefs in the United States and around the world.
MIT012	Identity and Difference	How can the individual be at once cause and consequence of society, a unique agent of social action and also a social product? This course explores how identities, whether of individuals or groups, based on single behaviors or institutional practices, are produced, maintained, and transformed. Students will be introduced to various theoretical perspectives that are used to make sense of identity formation, including essentialism, constructivism, stigma, deviance, discourse, and performance. We will explore the utility of these terms in discussing issues of gender, sexuality, race, ethnicity, religion, etc.
MIT013	Introduction to Anthropology	This class introduces students to the methods and perspectives of cultural anthropology. Readings emphasize case studies in very different settings (a nuclear weapons laboratory, a cattle-herding society of the Sudan, and a Jewish elder center in Los Angeles). Although some of the results and conclusions of anthropology will be discussed, emphasis will be on appreciating cultural difference and its implications, studying cultures and societies through long-term fieldwork, and most of all, learning to think analytically about other people's lives and our own.

MIT

ID	Título	Descripción
MIT014	Law and Society	Law is a common and yet distinct aspect of everyday life in modern societies. This course examines the central features of law as a social institution and as a feature of popular culture. We will explore the nature of law as a set of social systems, central actors in the systems, legal reasoning, and the relationship of the legal form and reasoning to social change. The course emphasizes the relationship between the internal logic of legal devices and economic, political and social processes. Emphasis is placed upon developing a perspective which views law as a practical resource, a mechanism for handling the widest range of unspecified social issues, problems, and conflicts, and at the same time, as a set of shared representations and aspirations.
MIT015	Magic, Witchcraft, and the Spirit World	Spiritual, magical, and "occult" aspects of human behavior in anthropological and historical perspective: magic, ritual curing, trance, spirit possession, sorcery, and accusations of witchcraft. Material drawn from traditional nonwestern societies, medieval and early modern Europe, and colonial and contemporary North America.
MIT016	Marketing, Microchips and McDonalds: Debating Globalization	We will explore the range of experiences of law for its ministers (lawyers, judges, law enforcement agents and administrators) as well as for its supplicants (citizens, plaintiffs, defendants). We will examine how law is mobilized and deployed by professionals and ordinary citizens. We cannot cover all aspects of the legal system, nor focus on all the different actors. A set of topics has been selected to develop understanding of the situational and systemic demands within which actors in the legal system operate and perform their roles; at the same time, we will try to discover systematic patterns in the uses and consequences of law. Throughout the course there is concern for understanding what we mean by legality and the rule of law.
MIT017	Medical Anthropology: Culture, Society, and Ethics in Disease and Health	This course looks at medicine from a cross-cultural perspective, focusing on the human, as opposed to biological, side of things. Students learn how to analyze various kinds of medical practice as cultural systems. Particular emphasis is placed on Western (bio-) medicine; students examine how biomedicine constructs disease, health, body, and mind, and how it articulates with other institutions, national and international.
MIT018	Medicine, Religion and Politics in Africa and the African Diaspora	This course provides an exploration of colonial and postcolonial clashes between theories of healing and embodiment in the African world and those of western bio-medicine. It examines how Afro-Atlantic religious traditions have challenged western conceptions of illness, healing, and the body and have also offered alternative notions of morality, rationality, kinship, gender, and sexuality. It also analyzes whether contemporary western bio-medical interventions reinforce colonial or imperial power in the effort to promote global health in Africa and the African diaspora.

MIT

ID	Título	Descripción
MIT019	Myth, Ritual, and Symbolism	Human beings are symbol-making as well as tool-making animals. We understand our world and shape our lives in large part by assigning meanings to objects, beings, and persons; by connecting things together in symbolic patterns; and by creating elaborate forms of symbolic action and narrative. In this introductory subject we consider how symbols are created and structured; how they draw on and give meaning to different domains of the human world; how they are woven into politics, family life, and the life cycle; and how we can interpret them.
MIT020	Photography and Truth	The semester will be devoted to a number of topics in symbolism.
MIT021	Race and Science	Metaphor and Other Figurative Language
MIT022	Seminar in Ethnography and Fieldwork	The Raw Materials of Symbolism, especially Animals and The Human Body
MIT023	Social Theory and Analysis	Cosmology and Complex Symbolic Systems
MIT024	Technology and Culture	Ritual, including Symbolic Curing and Magic
MIT025	The Anthropology of Biology	Narrative and Life
MIT026	The Anthropology of Computing	Mythology
MIT027	The Anthropology of Cybercultures	This course explores a range of contemporary scholarship oriented to the study of 'cybercultures,' with a focus on research inspired by ethnographic and more broadly anthropological perspectives. Taking anthropology as a resource for cultural critique, the course will be organized through a set of readings chosen to illustrate central topics concerning the cultural and material practices that comprise digital technologies. We'll examine social histories of automata and automation; the trope of the 'cyber' and its origins in the emergence of cybernetics during the last century; cybergeographies and politics; robots, agents and humanlike machines; bioinformatics and artificial life; online sociality and the cyborg imaginary; ubiquitous and mobile computing; ethnographies of research and development; and geeks, gamers and hacktivists. We'll close by considering the implications for all of these topics of emerging reconceptualizations of sociomaterial relations, informed by feminist science and technology studies.
MIT028	The Anthropology of Sound	This class examines the ways humans experience the realm of sound and how perceptions and technologies of sound emerge from cultural, economic, and historical worlds. In addition to learning about how environmental, linguistic, and musical sounds are construed cross-culturally, students learn about the rise of telephony, architectural acoustics, and sound recording, as well as about the globalized travel of these technologies. Questions of ownership, property, authorship, and copyright in the age of digital file sharing are also addressed. A major concern will be with how the sound/noise boundary has been imagined, created, and modeled across diverse sociocultural and scientific contexts. Auditory examples — sound art, environmental recordings, music — will be provided and invited throughout the term.

MIT

ID	Título	Descripción
MIT029	The Conquest of America	In this course the conquest and colonization of the Americas is considered, with special attention to the struggles of native peoples in Guatemala, Canada, Brazil, Panama, and colonial New England. In two segments of the course-one devoted to the Jesuit missionization of the Huron in the 1630s, the other to struggles between the government of Panama and the Kuna between 1900 and 1925-students examine primary documents such as letters, reports, and court records, to draw their own conclusions. Attention focuses on how we know about and represent past eras and other peoples, as well as on the history of struggles between native Americans and Europeans.
MIT030	The Contemporary American Family	We begin by considering briefly the evolution of the family, its cross-cultural variability, and its history in the West. We next examine how the family is currently defined in the U.S., discussing different views about what families should look like. Class and ethnic variability and the effects of changing gender roles are discussed in this section. We next look at sexuality, traditional and non-traditional marriage, parenting, divorce, family violence, family economics, poverty, and family policy. Controversial issues dealt with include day care, welfare policy, and the "Family Values" debate.
MIT031	Violence, Human Rights, and Justice	This course examines the contemporary problem of political violence and the way that human rights have been conceived as a means to protect and promote freedom, peace and justice for citizens against the abuses of the state.

Tabla 8-6 Categoría de antropología en repositorio EduTube.

EDUTUBE			
ID	**Titulo**	**Descripción General**	**Tags**
EDU001	Chimamanda Adichie: The danger of a single story (TEDTalks)	A powerful, inspirational - and at times humorous - lecture by acclaimed Nigerian writer Chimamanda Adichie about identity, prejudice, misunderstanding, and the danger of single stories.	TEDtalks Prejudice Nigeria Chimamanda Adichie Africa
EDU002	China	Fotos de China	China
EDU003	Cold Cuts - Cuisine of the Arctic	A documentary about the culinary culture of the Inuit people which is high in protein and fat, and their tradition of hunting caribou, seal, walrus, polar bears, muskoxen and whales.	Greenland Raw meat Inuit Arctic
EDU004	Did Aliens build the Pyramids? (1of3)	Highlights the pyramids and other inventions that illustrate the brilliance of ancient man - are they so brilliant that they could only have been built by aliens?	
EDU005	Did Aliens build the Pyramids? (2of3)	Part 2 of "Did aliens build the pyramids?"	
EDU006	Did Aliens build the Pyramids? (3of3)	Part 3 of "Did Aliens build the Pyramids?"	
EDU007	Documentaire West Afrika (Togo)		
EDU008	Elongated Skulls Discovered in Russia	Elongated skulls similar to those found in South America have been discovered in Russia. Altering the shape of the head through head binding was also practiced by ancient Egyptians, Australian Aborigines and certain tribes of North American natives.	
EDU009	How Australian Aboriginals created fire	An elder of an Australian aboriginal tribe demonstrates how to create "nature's secret fire" using sticks of wood. Fire was important not just for cooking, but also to stimulate growth of certain plant foods	Fire Bushfires Australian aboriginals Australia
EDU010	Inuit Kiss	An "Eskimo kiss" is more than just rubbing noses, as this video demonstrates.	Inuit Arctic
EDU011	Inuit Throat Singing	Watch the unique and fascinating musical expression of Inuit throat singing - a mixture of husky chanting and low growling.	singing Music explore.org Arctic

EDUTUBE

ID	Titulo	Descripción General	Tags
EDU012	Kumu Kahua Theatre - Hawaii	The Kumu Kahua Theatre in Honolulu keeps alive the stories and culture of Hawaii through plays for, by and about the people of Hawaii.	theatre Performing Arts Hawaii explore.org Art Acting
EDU013	Life in the Wilderness	A fascinating insight into life in the New Zealand bush, and what it is like to have lived there all your life - no cars, no schools, no TVs or computers, and sometimes without seeing anyone else for months.	
EDU014	Maasai Culture and Dance	This short video provides a glimpse of the rich Maasai culture and their energetic music and dance.	Tanzania Maasai Kenya explore.org East Africa Culture anthropology Africa
EDU015	Petra		Petra Middle East Jordan explore.org archeology
EDU016	Racism in European football	Racism in football is a serious worldwide problem, and it is not only fans who are guilty of racism. The video includes an interview with French star footballer Thierry Henrey (twice nominated FIFA World Player of the Year), who is an active spokesperson against racism in football.	Discrimination Thierry Henry Racism Football
EDU017	Snake Charmer	Harihar Rao discusses the inner workings of the snake charmer instrument.	snake charmer India explore.org Culture Animals
EDU018	Steven Pinker: A brief history of violence	Contrary to popular belief, our ancestors were far more violent than we are. Today we probably live in the most peaceful time of our species' existence.	Violence
EDU019	The Mayan Calendar	Maya civilization is noted for its fully developed written language, art, architecture, mathematical, astronomical systems as well as its calendars. The Maya Long Count calendar ends on December 21, 2012, and this day is therefore considered by many to be significant.	

Tabla 8-7 Categoría de antropología en repositorio Merlot.

MERLOT		
ID	**Título**	**Descripción**
MER001	Virtual Instrument Museum	This site provides a "virtual museum" of instruments from various parts of the world, including North
MER002	Edo Japan, A Virtual Tour	A complex and highly detailed virtual tour of 18th century Edo (now Tokyo) using traditional Japanese
MER003	The Internet Sacred Text Archive	Probably the most extensive single site on the Internet for sacred texts from all sorts of religions
MER004	Becoming Human	Una web multimedia la evolución humana. Un viaje en el tiempo y la prehistoria para ver los orígenes de la
MER005	Ayiti: The Cost of Life	In this game developed for Unicef, players guide a Haitian family of five through their struggle to..
MER006	Nonverbal Behavior / Nonverbal Communication. Links.	Very comprehensive collection of links to online resources on nonverbal behavior: papers and abstracts,...
MER007	Science and Race: Concept and Category	What is race? Does everyone think about race in the same way? How did the concept of "race" evolve?
MER008	Kinship and Social Organization	This colorful, sometimes animated tutorial presents the principles of kinship, marriage, and residence. ...
MER009	Utah State University Open Courseware (OCW)	USU OCW is a free and open educational resource for faculty, students, and self-learners, throughout Utah...
MER010	The Indivisible: Stories of American Community	Educator's Guide, examines a national documentary project about twelve diverse communities exploring the...
MER011	Anthropology Links (U of Pavia)	A broad collection of anthropology links maintained by the University of Pavia
MER012	Looking into the Westside: Untold Stories of the People, 1900-1997	This University of Arizona web exhibit focuses on eight youth historians from Tucson's Westside
MER013	African Mudcloth	Investigates the making of mudcloth. Includes links to the following: a brief history, map of Mali, an...
MER014	Lake Titicaca PowerPoint Matching Exercise	This PowerPoint activity is part of the unit on "Las Islas Uros" that I've posted here. In it, students...
MER015	Reframing America: Photography through the Eyes of Immigrants	Reframing America explores aspects of immigration revealed through photographs taken by immigrant...
MER016	Paris Codex	The Paris Codex is a digitally reproduced version of an ancient Maya book. Pre-Columbian Maya texts are...
MER017	Whose Mummy is it?	Whose Mummy Is It?? is a new way to learn about an old subject. It is a complete mini-course in...
MER018	Introduction to Paleoanthropology	Introduces the field of physical anthropology; Physical anthropology: study of human biology, nonhuman…

MERLOT		
ID	**Título**	**Descripción**
MER019	MIT Open Courseware (OCW) Collection	MIT OCW is a large scale, web-based electronic publishing initiative whose goals are to : Provide free,...
MER020	Astronomy of Many Cultures	This annotated listing includes over 90 books, articles, and websites that deal with the astronomy...
MER021	Virtual Autopsy: The Ice Maiden, Emissary to the Gods	This Natioanl Geographic site displays the remains of the Ice Maiden found during the Andes expedition.
MER022	Aaron Siskind and Max Yavno Archives Photographs of Mexico	Educator's Guide features photographs of Mexico and complements study in many ...
MER023	adherents.com	VERY extensive database of population and demographic statistics regarding thousands of religious groups
MER024	ArchaeoSim	Explores social and environmental tradeoffs in the farming civilization of ancient Subir (closely related...
MER025	Communal Living in Russia: A Virtual Museum	This Web site--an online ethnographic museum--explores and explains a striking social phenomenon
MER026	Defending and Attacking Polygamy in Saudi Arabia	Listen to this report by Julie McCarthy who looked at polygamy in Saudi Arabia. This audio can be used...
MER027	Las Islas Uros Lesson Plan	This document is the central lesson plan for the unit on "Las Islas Uros" that I've posted here.
MER028	Key for Lake Titicaca PowerPoint Matching Exercise	This is the key to the PowerPoint activity that is part of the unit on "Las Islas Uros" that I've posted...
MER029	Introduction to Australian Archaeology	In this introductory course Part 1, we consider what archaeology is, and how it fits into the scheme of...
MER030	Quipu: Dedicated to Researchers of Andean Anthropology	A large, well-maintained collection of Andean archeology research resources. Includes news, research,...
MER031	Nature Reserves in Jordan	Talks about the six nature reserves in Jordan: Ajloun, Azraq, Dana, Shaumari, Wadi Mujib, and Wadi Rum.
MER032	Um Qais or Gadara	It talks about the history of Um Qais, Jordan and what you can see there.
MER033	Regarding Diversity	Through documentary photography and interviews we explore the social repercussions of cultural diversity...
MER034	Les Classiques des sciences sociales	Une bibliotheque virtuelle d'ouvrages fondamentaux en sociologie, anthropologie, economie politique
MER035	Tea and Zen	This website provides basic information about the relationship between the ritual preparation of tea and zen.
MER036	Las Islas Uros Video	This video consists of a 38 minute long Spanish-language presentation about life on the Floating Uros.
MER037	Una Casa en Las Islas Uros Video	This 14-minute-long video interview with a woman resident of the Uros Islands is a part of the unit..

MERLOT		
ID	**Título**	**Descripción**
MER038	SurLaLune Fairy Tales	SurLaLune Fairy Tales features 49 annotated fairy tales, including their histories, similar tales across..
MER039	Directories/Indexes for Determining Publisher Open Access Status	Romeo lists the status of publisher copyright policies and author-archiving policies of academic...
MER040	Archaeology Online	India is indeed a place of great antiquity and great mystery - the culture is deeply mysterious, but by.
MER041	Don Antonio Zepeda: A Story of Four Generations	This University of Arizona web exhibit tells the story of family members of the Zepeda family.
MER042	Darwin Test One	An interactive self-test based on a selected reading which is an overview of Darwin's life, this website...
MER043	mediatedcultures.net	This site is home of the digital ethnography working group, a team of cultural anthropology...
MER044	Cultural Anthropology	This course provides a solid introduction for students who are new to the branch of cultural...
MER045	Folklife and Fieldwork A Layman's Introduction to Field Techniques	
MER046	Cultural Anthropology/Human Rights	-
MER047	Life in the Palaeozoic	This is a free, online textbook/course that starts "by looking at the Cambrian explosion, when many forms..
MER048	Visual Anthropology	This module is concerned with how anthropology can contribute to - and gain insight from - the analysis...
MER049	Anthropological ideas	Anthropological Ideas introduces the key ideas and perspectives that will enable students to complete...
MER050	The Hashemite Kingdom of Jordan	It's a travel guide through Jordan.
MER051	The Chaco Digital Initiative	The Chaco Digital Initiativeis a collaborative effort to create a digital archive that will integrate...
MER052	Religion & Social Order	This course is an anthropological exploration of religions in diverse cultural and historical contexts....
MER053	Exploring religions and cultures	This course is an introduction to anthropology through an exploration of western and non-western cultures...
MER054	ArchaeologyInfo.com	An online museum with pictures and articles about human ancestors, along with links to books and other...
MER055	Elixr: Student Video Projects in Anthropology Education	-
MER056	American Radioworks	AMERICAN RADIOWORKS® is the national documentary unit of American Public Media. It is public radio's...
MER057	Diverse Contexts of Human Infancy	-

MERLOT		
ID	**Título**	**Descripción**
MER058	Enduring Legacies Native Cases	These resources are contemporary teaching case studies related to Native Americans and their lived...
MER059	The Secret in the Cellar	The Secret in the Cellar webcomic is a forensic science mystery based closely on the case of the Leavy...
MER060	Peace: Contemplating Contentment	Using lessons from various religions around the world based on the lessons of self-examination they teach...
MER061	Latin American & Latino Studies Reader	The reader consists of 20 texts that enhance students' reading skills and knowledge of Latin American and…
MER062	Global Development Network (GDN) - Free Journal Access Portal	GDN has linked policy research institutes from 11 regions and more than 100 countries. GDN offers a range...

Tabla 8-8 Categoría de antropología en repositorio Connexions.

CONNEXIONS			
ID	**Título**	**Descripción**	**Tags**
CON001	Free Online Anthropology Videos and Video Clips	A list of online videos and video clips for anthropology or cross cultural classes. These can be used for online classes, hybrids, classroom presentations, or outside	anthropology, cross-cultural, culture, diversity, ethnographies, online, psychology, videos
CON002	Cultural Anthropology Ethnography Assignment	This is a mini-ethnography assignment used for a major class project in Cultural Anthropology. Sections include: selection of an informant, background research, interview protocol and interview questions, paper requirements, grading criteria, and sample paper.	anthropology, assignment, ethnography, interview, project
CON003	Negro and White Exclusion Towns in Indian Territory and Oklahoma		This module is a republication of the following essay: Frank G. Speck. 1907. Negro and White Exclusion Towns in Indian Territory. Southern Workman 36, no. 8: 430-432. Based on ethnographic field research undertaken in Oklahoma and Indian Territories in 1904 and 1905, Speck's essay describes the racial polarization ... this basis.

CONNEXIONS

ID	Título	Descripción	Tags
CON004	Notes on Creek Mythology (	This module is a republication of the following essay: Speck, Frank G. 1909. Notes on Creek Mythology. Southern Workman 38, no. 1: 9-11. Based on ethnographic field research undertaken in the Creek Nation in 1904, 1905, and 1908, Speck's essay describes the major tale types and motifs characterizing ... this basis.	African American, American Indian, Bungling Host, Catchword, Comparative Folklore, Creation, Creek Nation, Cultural Anthropology, Culture, Culture Hero, Earth Diver, Ethnography, Ethnology, Folklore, Folkloristics, Folktale, Hare and Tortoise, Indian Territory, Legend, Magical Flight, Motif, Muscogee, Muscogee (Creek) Nation, Mythology, Native American, Oklahoma, Rabbit Story, Stone Clad, Tale Type, Tar Baby, Theft of Fire, Trickster, Verbal Art
CON005	Title Page	This is the title page for the collection Negro and White Exclusion Towns and Other Observations in Oklahoma and Indian Territory: Essays by Frank G. Speck from The Southern Workman	Anthropology, Cultural Geography, Cultural History, Ethnography, Ethnology, Social Geography, Social History
CON006	The "Savage Mind" on Madison Avenue: A Structural Analysis of Television Advertising	n analysis of television advertising video recordings from the 1970's using a methodology based on the structural anthropology of Claude Lévi-Strauss.	Levi-Strauss, Media Ecology, Media History, Media Studies, Structural Anthropology
CON007	Editor's Introduction: On Frank G. Speck's Oklahoma and Indian Territory Essays for the Southern Workman	This module is the editor's introduction to the collection of essays by Frank G. Speck published under the title Negro and White Exclusion Towns and Other Observations in Oklahoma and Indian Territory.	Anthropology, Cultural Geography, Cultural History, Ethnology, Folklore, Freedmen, History, Indian Territory, Missionization, Muscogee (Creek), Oklahoma, Race Reltions, Racism, Social History, Violence
CON008	An Introduction to the Class (Text as Property/Property as Text)	An introduction to the Rice University course "Text as Property/Property as Text" which seeks to compare ancient and modern conceptions of authorship, ownership and alternative traditions of writing, stewardship, allusion, and distribution.	aesthetics, anthropology, classics, comparison, cultural analysis, economic value, intellectual property law, law, literature

CONNEXIONS

ID	Título	Descripción	Tags
CON009	Missions in the Creek Nation	This module is a republication of the following essay: Frank G. Speck. 1911. Missions in the Creek Nation. Southern Workman 40, no. 4: 206-208. Based on ethnographic field research undertaken in the Creek Nation in 1904, 1905 and 1908, Speck's essay describes the history and consequences of Christian ... this basis.	African American, American Indian, Baptist, Christianity, Creek Nation, Cultural Anthropology, Culture, Ethnography, Ethnology, Euchee, European American, History, Indian Territory, Methodist, Mission, Missionaries, Missionization, Muscogee, Muscogee (Creek) Nation, Native American, Oklahoma, Social Conditions, Social Organization, Social Problems, Society, Yuchi
CON010	The Negroes and the Creek Nation	This module is a republication of the following essay: Frank G. Speck. 1908. The Negroes and the Creek Nation. Southern Workman 37, no. 2: 106-110. Based on ethnographic field research undertaken in the Creek Nation, Indian Territory in 1904 and 1905, Speck's essay describes the history and present ... this basis.	African American, American Indian, Creek Nation, Cultural Anthropology, Culture, Ethnography, Ethnology, European American, History, Indian Territory, Muscogee, Muscogee (Creek) Nation, Native American, Oklahoma, Race, Racism, Social Conditions, Social Organization, Social Problems, Society
CON011	Observations in Oklahoma and Indian Territory	This module is a republication of the following essay: Frank G. Speck. 1907. Observations in Oklahoma and Indian Territory. Southern Workman 36, no. 1: 23-27. Based on ethnographic field research undertaken in the Oklahoma and Indian Territories in 1904 and 1905, Speck's essay describes a range of environmental ... this basis.	Cultural Anthropology, Cultural Geography, Culture, Ethnography, Ethnology, History, Indian Territory, Oklahoma, Oklahoma Territory, Social Conditions, Social Organization, Social Problems, Society
CON012	Tips for Reading	This module contains suggestions for how to read a variety of texts generally, with special emphasis on issues of authorship, ownership and the historical and legal context.	classics, cultural analysis, reading practices, Tips
CON013	What does 'shaman' mean? The dispute surrounding the definition and correct use of the word 'shaman'	This module includes a short article (1 1/2 pgs) and suggested lesson plan on the dispute among anthropologists surrounding the definition and uses of the word 'shaman.' I think that it would be most appropriate for a high school social studies class, or, of course, for anyone who is just curious about the subject.	anthropologist, anthropology, lesson plan grades 8-12, shaman, shamanism, social studies

CONNEXIONS

ID	Título	Descripción	Tags
CON014	Cholera, Canker Rash and Consumption: historical epidemiology and nosology in Massachusetts, 1850-1920	"Cholera, Canker Rash and Consumption: historical epidemiology and nosology in Massachusetts, 1850-1920", A recorded Public Health Seminar delivered by Professor Alan C. Swedlund on November 9, 2009. Alan C. Swedlund is Professor Emeritus in the Department of Anthropology at University of Massachusetts, Amherst.	Canker, Cholera, Epidemiology, Health, Public, Rash
CON015	Gang Redux: A Balanced Anti-Gang Strategy	Department of Education Brownbag Lecture "Gang Redux: A Balanced Anti-Gang Strategy" by Professor James Diego Vigil recorded on Monday, May 17, 2010. James Diego Vigil, Ph.D., professor in the Department of Criminology, Law and Society, University of California, Irvine. Professor Vigil holds a Ph.D. and an M.A. in Anthropology from the University of California, Los Angeles.	Anti-gang, Criminology, Gangs, Incarceration
CON016	Mitali Banerjee	Biography of Mitali Banerjee. Anthropology & Electrical engineering, Class of 2004, Rice University.	Banerjee, Mitali
CON017	Africa: Beginning to 8000 B.C.		Africa, Delineation, History, Maxfield, World
CON018	America: 0 to 100 A.D.		America, History, Maxfield, World
CON019	America: 200 to 101 B.C.		America, History, Maxfield, World
CON020	America: A.D. 101 to 200		America, History, Maxfield, World
CON021	Ch. 3 British Colonial America (1588-1701)	his chapter examines early British settlements in North America roughly from 1588 to 1701.	Act of Toleration, Bacon's Rebellion, Barbados Slave Code of 1661, Dominion of New England, Elizabeth I, Free Booters, Indentured servants, Jamestown, John Rolfe, John Smith, Joint Stock Company, King Phillipp's War, New England Confederation, Powhatten Wars, Proprietary colony, Puritans, Royal Charter, Sir Edmund Andros, Sir Francis Drake, Spanish Armada

CONNEXIONS			
ID	**Título**	**Descripción**	**Tags**
CON022	The Challenge of Integrating Democratic Community, Social Justice, and School Improvement in Educational Leadership Programs (m34643)	This article proposes, as Joseph Murphy did over a decade ago, that educational leadership preparation programs integrate democratic community, social justice, and school improvement and promote these three principles as the central foci of educational leadership. First, the article discusses why each of the three principles should be emphasized in ... program development.	democratic community, educational leadership, social justice, university preparation
CON023	Conference Participants		archive, archives, archiving, Digital, Humanities, librarianship, libraries, library, Online, Pedagogy, Scholarship, University
CON024	Creating the Eduerati: Professorial Leadership to Create K-12 Educational System Change	Our decentralized education system originally built for a localized agrarian community now confronts a confluence of a highly networked and currently unstable global economy, instantaneous communication systems, vast wealth inequities, unstoppable human migrations, impending ecological disasters, ethnic/ religious violence and questionable ethnic conclusions. Can human intelligence and adaptability be summoned ... of thinking.	

8.6 Anexo 6 – Configuración del Archivo 'otech_crawl.txt' - Enlaces Web de la Categoría Antropología de los Repositorios

- http://www.edutube.org/en/taxonomy/term/7/feed
- http://www.edutube.org/en/category/anthropology
- http://www.edutube.org/en/category/anthropology?page=1
- http://ocw.mit.edu/rss/all/mit-allcourses-21A.xml
- http://www.merlot.org/merlot/materials.xml?category=2788&materialType=&keywords=&qstringrss=category%3D2788%26sort.property%3DoverallRating&sort.property=overallRating&sortbutton.x=18&sortbutton.y=7&sortbutton=Sort
- http://www.merlot.org/merlot/materials.htm?category=2788&&materialType=&keywords=&&qstringrss=category%3D2788%26sort.property%3DoverallRating&sort.property=overallRating&sortbutton.x=18&sortbutton.y=7&sortbutton=Sort
- http://www.merlot.org/merlot/materials.htm?pageSize=&page=2&category=2788&materialType=&keywords=&qstringrss=category%3D2788%26sort.property%3DoverallRating&sort.property=overallRating&sortbutton.x=18&sortbutton.y=7&sortbutton=Sort
- http://www.merlot.org/merlot/materials.htm?pageSize=&page=3&&category=2788&materialType=&keywords=&qstringrss=category%3D2788%26sort.property%3DoverallRating&sort.property=overallRating&sortbutton.x=18&sortbutton.y=7&sortbutton=Sort
- http://www.merlot.org/merlot/materials.htm?pageSize=&page=4&&category=2788&materialType=&keywords=&qstringrss=category%3D2788%26sort.property%3DoverallRating&sort.property=overallRating&sortbutton.x=18&sortbutton.y=7&sortbutton=Sort
- http://www.merlot.org/merlot/materials.htm?pageSize=&page=5&&category=2788&materialType=&keywords=&qstringrss=category%3D2788%26sort.property%3DoverallRating&sort.property=overallRating&sortbutton.x=18&sortbutton.y=7&sortbutton=Sort
- http://www.merlot.org/merlot/materials.htm?pageSize=&page=6&&category=2788&materialType=&keywords=&qstringrss=category%3D2788%26sort.property%3DoverallRating&sort.property=overallRating&sortbutton.x=18&sortbutton.y=7&sortbutton=Sort
- http://www.merlot.org/merlot/materials.htm?pageSize=&page=7&&category=2788&materialType=&keywords=&qstringrss=category%3D2788%26sort.property%3DoverallRating&sort.property=overallRating&sortbutton.x=18&sortbutton.y=7&sortbutton=Sort
- http://cnx.org/lenses/cnxhcc/affiliation/atom
- http://cnx.org/content/search?sorton=weight&view_mode=detail&words=Anthropology&template=/content/search&b_size=25&allterms=weakAND

8.7 Anexo 7 - Configuración del Archivo 'crawl-urlfilter.txt' de Nutch.

- +^http://www.edutube.org/en/taxonomy/term/7/feed
- +^http://www.edutube.org/en/category/anthropology
- +^http://www.edutube.org/en/category/anthropology?page=1
- +^http://ocw.mit.edu/rss/all/mit-allcourses-21A.xml
- +^http://www.merlot.org/merlot/materials.xml?category=2788&materialType=&keywords=&qstringrss=category%3D2788%26sort.property%3DoverallRating&sort.property=overallRating&sortbutton.x=18&sortbutton.y=7&sortbutton=Sort
- +^http://www.merlot.org/merlot/materials.htm?category=2788&&materialType=&keywords=&&qstringrss=category%3D2788%26sort.property%3DoverallRating&sort.property=overallRating&sortbutton.x=18&sortbutton.y=7&sortbutton=Sort
- +^http://www.merlot.org/merlot/materials.htm?pageSize=&page=2&category=2788&materialType=&keywords=&qstringrss=category%3D2788%26sort.property%3DoverallRating&sort.property=overallRating&sortbutton.x=18&sortbutton.y=7&sortbutton=Sort
- +^http://www.merlot.org/merlot/materials.htm?pageSize=&page=3&&category=2788&materialType=&keywords=&qstringrss=category%3D2788%26sort.property%3DoverallRating&sort.property=overallRating&sortbutton.x=18&sortbutton.y=7&sortbutton=Sort
- +^http://www.merlot.org/merlot/materials.htm?pageSize=&page=4&&category=2788&materialType=&keywords=&qstringrss=category%3D2788%26sort.property%3DoverallRating&sort.property=overallRating&sortbutton.x=18&sortbutton.y=7&sortbutton=Sort
- +^http://www.merlot.org/merlot/materials.htm?pageSize=&page=5&&category=2788&materialType=&keywords=&qstringrss=category%3D2788%26sort.property%3DoverallRating&sort.property=overallRating&sortbutton.x=18&sortbutton.y=7&sortbutton=Sort
- +^http://www.merlot.org/merlot/materials.htm?pageSize=&page=6&&category=2788&materialType=&keywords=&qstringrss=category%3D2788%26sort.property%3DoverallRating&sort.property=overallRating&sortbutton.x=18&sortbutton.y=7&sortbutton=Sort
- +^http://www.merlot.org/merlot/materials.htm?pageSize=&page=7&&category=2788&materialType=&keywords=&qstringrss=category%3D2788%26sort.property%3DoverallRating&sort.property=overallRating&sortbutton.x=18&sortbutton.y=7&sortbutton=Sort
- +^http://cnx.org/lenses/cnxhcc/affiliation/atom
- +^http://cnx.org/content/search?sorton=weight&view_mode=detail&words=Anthropology&template=/content/search&b_size=25&allterms=weakAND

Saltar en Edutube

- -http://www.edutube.org/category/biology
- -http://www.edutube.org/category/animals--wildlife
- -http://www.edutube.org/category/biology
- -http://www.edutube.org/category/chemistry
- -http://www.edutube.org/category/education
- -http://www.edutube.org/tags/guitar-lessons
- -http://www.edutube.org/category/math-statistics
- -http://www.edutube.org/category/space-astronomy
- -http://www.edutube.org/category/technology
- -http://www.edutube.org/sitemap
- -http://www.edutube.org/en/search_video
- -http://www.edutube.org/en/node/add
- -http://www.edutube.org/en/node/add/podcasts
- -http://www.edutube.org/en/node/add/video
- -http://www.edutube.org/en/%252FAbout-edutube
- -http://www.edutube.org/en/contributors
- -http://www.edutube.org/en/edutube-faq
- -http://www.edutube.org/en/forum
- -http://www.edutube.org/en/groups
- -http://www.edutube.org/en/rules-and-policies
- -http://www.edutube.org/en/user/register
- -http://www.edutube.org/en/user/password
- -http://www.edutube.org/en/category/anthropology?sort=asc&order=Length&page=1
- -http://www.edutube.org/en/category/anthropology?sort=asc&order=Added&page=1
- -http://www.edutube.org/en/category/anthropology?sort=asc&order=Views%2Fday&page=1
- -http://www.edutube.org/en/category/anthropology?sort=asc&order=EduTube+Index&page=1

#Saltar Merlot

- -^http://([a-z0-9]*\.)*merlot.org/merlot/portfolios.htm
- -^http://([a-z0-9]*\.)*merlot.org/merlot/reviews.htm
- -http://www.merlot.org/merlot/index.htm
- -http://www.merlot.org/merlot/communities.htm
- -http://www.merlot.org/merlot/members.htm?sort.property=contributions
- -http://www.merlot.org/merlot/login.htm
- -http://taste.merlot.org/
- -http://www.merlot.org/merlot/join.htm
- -http://www.merlot.org/merlot/login.htm?page=none
- -http://www.merlot.org/merlot/materials.htm?category=2789&materialType=&keywords=&&qstringrss=category%3D2788%26sort.property%3DoverallRating&sort.property=overallRating&sortbutton.x=18&sortbutton.y=7&sortbutton=Sort
- -http://www.merlot.org/merlot/materials.htm?category=2802&materialType=&keywords=&&qstringrss=category%3D2788%26sort.property%3DoverallRating&sort.property=overallRating&sortbutton.x=18&sortbutton.y=7&sortbutton=Sort
- -http://www.merlot.org/merlot/materials.htm?category=2803&materialType=&keywords=&&qstringrss=category%3D2788%26sort.property%3DoverallRating&sort.property=overallRating&sortbutton.x=18&sortbutton.y=7&sortbutton=Sort
- -http://www.merlot.org/merlot/materials.htm?category=2804&materialType=&keywords=&&qstringrss=category%3D2788%26sort.property%3DoverallRating&sort.property=overallRating&sortbutton.x=18&sortbutton.y=7&sortbutton=Sort
- -http://www.merlot.org/merlot/materials.htm?category=2805&materialType=&keywords=&&qstringrss=category%3D2788%26sort.property%3DoverallRating&sort.property=overallRating&sortbutton.x=18&sortbutton.y=7&sortbutton=Sort
- -http://www.merlot.org/merlot/materials.htm?category=2806&materialType=&keywords=&&qstringrss=category%3D2788%26sort.property%3DoverallRating&sort.property=overallRating&sortbutton.x=18&sortbutton.y=7&sortbutton=Sort
- -http://www.merlot.org/merlot/materials.htm?category=2825&materialType=&keywords=

&&qstringrss=category%3D2788%26sort.property%3DoverallRating&sort.property=overallRating&sortbutton.x=18&sortbutton.y=7&sortbutton=Sort

- -http://www.merlot.org/merlot/materials.htm?category=2826&materialType=&keywords=&&qstringrss=category%3D2788%26sort.property%3DoverallRating&sort.property=overallRating&sortbutton.x=18&sortbutton.y=7&sortbutton=Sort
- -http://www.merlot.org/merlot/materials.htm?category=2827&materialType=&keywords=&&qstringrss=category%3D2788%26sort.property%3DoverallRating&sort.property=overallRating&sortbutton.x=18&sortbutton.y=7&sortbutton=Sort
- -http://www.merlot.org/merlot/materials.htm?category=2828&materialType=&keywords=&&qstringrss=category%3D2788%26sort.property%3DoverallRating&sort.property=overallRating&sortbutton.x=18&sortbutton.y=7&sortbutton=Sort
- -http://www.merlot.org/merlot/viewMember.htm?id=anonymous
- -http://taste.merlot.org/allpartnerlist.html
- -http://taste.merlot.org/faq.html
- -http://www.merlot.org/merlot/materials.htm?sort.property=overallRating

#Salatar en connexions

- -^http://cnx.org/
- -http://cnx.org/lenses
- -^http://cnx.org/aboutus/
- -^http://cnx.org/help/
- -^http://cnx.org/login_form?came_from=mydashboard
- -http://cnx.org/content/
- -^http://cnx.org/join_form
- -^http://cnx.org/mail_password_form
- -^http://cnx.org/content/col10151/latest/
- -^http://cnx.org/recentview_more
- -^http://cnx.org/lenses/vicarranz/vicarranz
- -^http://cnx.org/lenses/jenifferhomes/watch-americas-next-top-model-online-episodes
- -^http://cnx.org/lenses/derickfay/teaching-introductory-anthropology
- -^http://cnx.org/aboutus/people/sponsors

- -^http://www.hewlett.org/
- -^http://cnxconsortium.org/
- -^http://blog.cnx.org/

8.8 Anexo 8 – Ejemplo: Relación entre Exhaustividad y Precisión.

Para entender mejor el enunciado expresado anteriormente, se considerará el siguiente ejemplo explicado en (Martínez F. J., 2002): Supóngase que un usuario lleva a cabo una operación de recuperación de información en la cual inserta condiciones muy específicas, seguramente obtendrá un conjunto de resultados muy preciso pero, de igual modo, habrá dejado de recuperar algunos documentos a causa de ese alto nivel de especificación. Como ejemplo de esta situación, supóngase que se tienen estas dos operaciones de búsqueda que se plantean a continuación:

Consulta B1: “contaminación del agua en los ríos”
Consulta B2: “contaminación en los ríos”

Ambas búsquedas pretenden recuperar el mismo tipo de documento, pero, en el caso de la primera (B1), el usuario la plantea de una forma más específica que la segunda (B2). Este segundo usuario ha pensado que no es necesario emplear el término “agua” en la operación de recuperación de información, ya que, a lo mejor piensa que cuando se contamina un río, es el agua lo que se contamina y le ha parecido redundante e innecesaria tanta especificación.

Con toda seguridad, la primera búsqueda (B1) va a adolecer[38] del problema que estamos presentando, ya que basta que en un documento el autor o el indizador no haya alcanzado el nivel de especificación empleado por el usuario que plantea la búsqueda, para que sea recuperado por la segunda (B2) pero no por la primera (B1). En esta situación, la segunda búsqueda presentará unos niveles mayores de exhaustividad frente a la primera y unos niveles de precisión algo más bajos.

El caso contrario se presenta también frecuentemente: un usuario plantea una ecuación de búsqueda demasiado general, con la que seguramente recuperará la mayoría de los documentos relevantes con el tema de la cuestión, pero, al mismo tiempo recuperará muchos documentos que no resultan relevantes.

Esto implicará que los valores de precisión se reduzcan sustancialmente. Si por ejemplo, los usuarios del ejemplo anterior, hubieran realizado estas búsquedas:

[38] Estar sujeto a defectos. Disponible en: <http://www.definicion.org>

Consulta B1: “contaminación”
Consulta B2: “contaminación en los ríos”

La primera consulta (B1) obtendrá como resultado un gran número de documentos; es decir, el nivel de exhaustividad será mucho mayor, pero al mismo tiempo la precisión de los documentos recuperados se verá disminuida.

8.9 Anexo 9 – Encuesta al Usuario

8.9.1 Requerimientos de los SRI Propuestos

			Nutch			Regain		
	Requerimiento	**Descripción**	**Baja**	**Media**	**Alta**	**Baja**	**Media**	**Alta**
Interfaz de usuario	Amigabilidad	facilidad de acceso a las diversas funciones del SRI						
	Informatividad	capacidad de que el usuario pueda recibir información útil y no repetitiva de los registros consultados						
	Visualización de recuperación	todos los elementos visualizados sean oportunos, adecuados y se presenten de un modo agradable						
	Búsqueda avanzada	Búsqueda a mas detalle						
Req. no funcionales	Eficacia	el tiempo que toma el sistema o una parte del mismo para llevar a cabo una operación						
	Eficiencia	Consiste en que la información sea generada con el óptimo (más productivo y económico) uso de los recursos.						
	Integridad	La información sólo puede ser modificada por quien está autorizado y de manera controlada.						
	Disponibilidad	El SRI debe estar disponible cuando se le necesite.						
	Confiabilidad	se refiere a proporcionar la información apropiada						
	exhaustividad	habilidad del sistema para presentar todos los ítems relevantes						
	La precisión	o habilidad del sistema para presentar solamente ítems relevantes						
	Usabilidad	Que el sistema sea amigable						
Arquitectura	centralizada	Los componentes corren localmente en la máquina de búsqueda						
	distribuida	Los componentes pueden encontrarse en distintas áreas geográficas.						