UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA

La Universidad Católica de Loja

MODALIDAD CLÁSICA

ESCUELA DE CIENCIAS DE LA COMPUTACIÓN

DESARROLLO DE UN SISTEMA SEMIAUTOMÁTICO DE ETIQUETACIÓN DE CONTENIDOS DE BLOGS UTILIZANDO LENGUAJES DE DESCRIPCIÓN SEMÁNTICA.

Trabajo de fin de carrera previo a la obtención del título de Ingeniero en Sistemas Informáticos y Computación.

AUTORA

Burguán Valverde lliana Maritza

DIRECTORA

Ing. González Eras Alexandra Cristina

Loja – Ecuador

2011

CERTIFICACIÓN

Ing. Alexandra González

DIRECTORA DE TESIS

C E R T I F I C A:

Que el presente trabajo de investigación, previo a la obtención del título de INGENIERO EN SISTEMAS INFORMÁTICOS Y COMPUTACIÓN, ha sido dirigido, supervisado y revisado en todas sus partes, por lo mismo, cumple con los requisitos legales, exigidos, por la Universidad Técnica Particular de Loja, quedando autorizada su presentación.

Loja, junio del 2011

…….

Alexandra González Eras

DIRECTORA DE TESIS

CERTIFICACIÓN

Ing. Paola Sarango

CODIRECTORA DE TESIS

C E R T I F I C A:

Que el presente trabajo de investigación, previo a la obtención del título de INGENIERO EN SISTEMAS INFORMÁTICOS Y COMPUTACIÓN, ha sido dirigido, supervisado y revisado en todas sus partes, por lo mismo, cumple con los requisitos legales, exigidos, por la Universidad Técnica Particular de Loja, quedando autorizada su presentación.

Loja, junio del 2011

…….

Paola Sarango Lapo

DIRECTORA DE TESIS

AUTORÍA

El presente proyecto de tesis previa a la obtención del Título de Ingeniero en Sistemas Informáticos y Computación; sus conceptos, análisis y recomendaciones emitidas, es de absoluta responsabilidad de los autores.

Se debe indicar además que la información de otros autores en este trabajo está debidamente específica en fuentes de referencia y aparatos bibliográficos.

…….

Iliana Maritza Burguán Valverde

CESIÓN DE DERECHOS.

Yo, Iliana Maritza Burguán Valverde, 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 Art. 67 del Estatuto Orgánico de la Universidad Técnica Particular de Loja que su parte pertinente textualmente dice: ***"Forman parte del patrimonio de la Universidad de la propiedad intelectual de investigadores, 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"***

……..

Iliana Maritza Burguán Valverde

AGRADECIMIENTO

Agradezco a todas las personas que han contribuido en mi formación, tanto personal como profesional en especial a mis directoras de Tesis, Alexandra González y Paola Sarango, a demás del antiguo CITTES "Gestión del Conocimiento", ya que gracias a su orientación, motivación y paciencia, he podido culminar con éxito este trabajo.

.

Iliana

DEDICATORIA

Dedico esta tesis a Dios, que me ha permitido llegar hasta este punto consiguiendo alcanzar una de mis metas.

A mis padres Luz América y Edgar, quienes hicieron en mi una persona de bien, a demás de brindarme su apoyo a pesar de las situaciones que golpearon nuestras vidas. Todo lo que soy se los debo a ellos.

A mis hermanos Gino y Gildson, quienes siempre estuvieron conmigo apoyándome y aguantarme el stress que duró toda mi carrera.

Iliana

Tabla de contenido

Índice de Figuras

Indice de Tablas

RESUMEN

En este proyecto se desarrolla un sistema semiautomático de etiquetación de contenidos de blogs utilizando lenguajes de descripción semántica (Sistema de Etiquetación de Metadatos a Entradas de Blogs "EMEB"), en base a la extracción y reutilización de metadatos de las entradas publicados en el wordpress multiusuario de la Universidad Técnica Particular de Loja, a través de bookmarks estructurados en archivos RDF utilizando etiquetas Dublín Core, que pueden ser vistos en sistemas de representación de bookmarks ó por herramientas de representación semántica mediante el acceso a un directorio público. Además se aplica una taxonomía que permite la categorización de los blogs de acuerdo a la estructura organizacional de la universidad.

Cada uno de los bookmarks extraídos por el Sistema EMEB sirve como fuente de alimentación para el Repositorio del conocimiento KMS de la UTPL[1].

[1] Trabajo de fin de carrera previo a la obtención del título de Ingeniero en Sistemas Informáticos y Computación de la UTPL, realizado por Lorena del Cisne León Quiñonez. (http://www.utpl.edu.ec/repositorio/).

CAPITULO 1

1. SITUACIÓN ACTUAL DE EXTRACCION DE INFORMACIÓN SOBRE BLOGS

1.1 INTRODUCCIÓN

Las entradas que se publican en los blogs en la mayor parte de los casos no son etiquetados por sus autores, lo que provoca que tanto los motores de búsqueda tradicionales como especializados no ubiquen con precisión estos contenidos. Por tal razón las investigaciones en extracción de información toman amplio auge para conseguir una gestión eficaz y eficiente de la información.

En este capítulo se presenta un estado del arte de los sistemas de extracción de información, iniciando por la problemática que se desea resolver con el apoyo de los sistemas de recuperación de información.

1.2 PROBLEMÁTICA

El volumen de información que se genera en los blogs de la UTPL ha ido aumentado debido a que se ha constituido en un medio para la interacción académica de profesores y alumnos en las diferentes asignaturas, donde se publican trabajos de clase y proyectos que se desarrollan en el semestre. Pero en la actualidad sólo las personas que publicaron sus contenidos conocen de su existencia en la web, lo que limita el uso de esos recursos por otros usuarios.

Tomando como referencia el proyecto piloto "Un blog por Asignatura" (en el período Abril Agosto 2009), realizado por Gestión del Conocimiento en las Escuelas Comunicación Social y Electrónica y Telecomunicaciones.

Según la Tabla 1, donde se realizó la publicación de los trabajos realizados por los estudiantes en cada asignatura, se aprecia que no existe una cultura de etiquetación de contenidos, privando a los mismos de las etiquetas necesarias que faciliten una clasificación temática de contenidos de interés.

Tabla 1. Resumen de blogs académicos por asignatura de la Escuela de Comunicación Social

ESCUELA DE COMUNICACIÓN SOCIAL					
Asignaturas	**Entradas**	**Etiquetas**	**Entradas no Etiquetadas**	**Categorías**	**Entradas no Categorizadas**
Fotografía	4	6	1	4	1
Géneros periodísticos	19	6	10	6	10
Medios impresos	4	15	1	5	0
Marketing	28	15	25	3	18
Opinión pública	1	6	0	2	1
Periodismo digital	46	76	37	10	32
Administración de medios	2	5	1	2	2
Cine	2	2	2	2	1

Comunicación organizacional	28	30	17	10	12
Relaciones internacionales	60	60	28	4	26
Teoría de la imagen	2	7	0	3	1

En ésta Figura 1, se muestran algunas materias de la Escuela de Comunicación Social, en las que se evidencia el número de entradas de los blogs existen la mayoría de Entradas no etiquetadas, así como también las entradas no categorizadas.

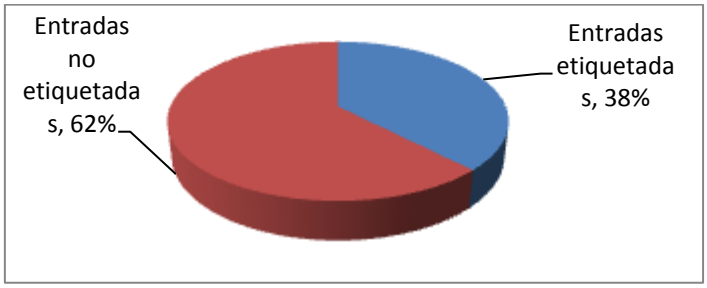


Figura 1. Entradas etiquetadas y entradas no etiquetadas en los blogs de materias de Comunicación Social según el proyecto de un blog por asignatura

Según la realidad en porcentajes de la Figura 1, se muestra que las entradas no etiquetadas es la gran mayoría el 62% del total de las entradas de las materias de la Escuela de Comunicación Social, mientras que si existen entradas etiquetadas en un 38%. Es claro que, cuando el usuario crea una entrada se olvida de colocar etiquetas o a su vez desconoce de esa característica que posee la plataforma wordpress.

Y según a la escuela de Electrónica y Telecomunicaciones en la Tabla 2, es preciso aclarar que durante el período Abril Agosto 2009, las entradas no etiquetadas son menores al total de entradas.

Tabla 2. Resumen de blogs académicos por asignatura de la Escuela de Electrónica y Telecomunicaciones

ESCUELA DE ELECTRÓNICA					
Asignaturas	**Entradas**	**Etiquetas**	**Entradas no Etiquetadas**	**Categorías**	**Entradas no Categorizadas**
Radiocomunicaciones	52	30	1	8	0
Sistemas de Información Geográfica	7	3	5	2	0
Computación de Altas Prestaciones	9	2	8	5	0
Circuitos y Sistemas	3	2	2	2	2
Ciencias Fundamentales	2	5	0	1	1
Revista Cortocircuito	10	29		2	

Según la Figura 2, se muestra la mayoría de las entradas son etiquetadas por usuarios y constan con un 81% de en el periodo Abril Agosto 2009, y con un 19% de las entradas etiquetadas.

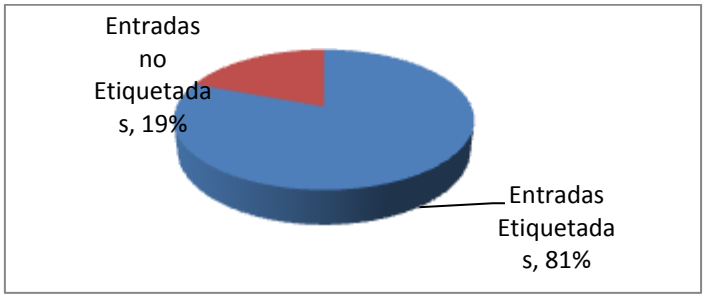


Figura 2. Blogs por asignaturas de la Escuela de Electrónica y Telecomunicaciones de las entradas etiquetadas y entradas no etiquetadas

En esta demostración del proyecto un Blog por Asignatura, se deduce que difiere el uso de los usuarios, ya que en la Figura 1, existe la mayoría de entradas sin etiquetar, en cambio en la Figura 2, es todo lo contrario, pero en tal razón, siguen siendo las entradas sin etiquetas donde se pierde valiosa información que aporta al conocimiento a la sociedad bloguera, lo cual da la pauta para que existan blogs a nivel de instituciones y es mas en la educación.

1.3 IMPORTANCIA DE LOS BLOGS EN LA EDUCACIÓN

En la Educación Superior los blogs se constituyen una bitácora por su flexibilidad para documentar las tareas previstas por el plan de estudios, a través de los comentarios se pueden promover debates que enriquecen el proceso educativo, facilitan el seguimiento de tareas y progresos del alumno convirtiéndose en un canal de comunicación o tutorías.

Dentro del modelo educativo universitario "Los blogs pueden ser una herramienta suficiente para la investigación del profesor, particularmente como archivo de fuentes y datos empíricos, y la comunicación con otros colegas investigadores"[1], siendo una de las interacciones más atractivas desde el punto de vista educativo, y sobre todo, desde una visión constructiva del aprendizaje.

La plataforma usada para este tipo de blogs educativos es WordPress, según [2]. Existen tanto Wordpress independientes y WPMU[2], pero siendo el segundo como el más utilizado en organizaciones educativas.

El Wordpress Multiusuario permite crear un sitio proveedor de blogs, del cual otras personas se registran para crear sus propios blogs o participar de blogs ya creados.

La implementación del Wordpress multiusuario en la Universidad Técnica Particular de Loja se debió a las características que presenta como:

- ✓ Facilidad en la administración de blogs por parte de usuarios y webmaster.
- ✓ Estandarización de interfaces de blog.
- ✓ Alojamiento de un dominio institucional

Los blogs creados con el WPMU son como cualquier otro blog, constan de entradas (distribuidas por lo general en categorías), comentarios y widgets.

[2] WordPress Multiusuario

La página principal de un blog es la página de Inicio, donde se van a ver las entradas o posts que el autor crea.

Por último, un blog tiene widgets. Los Widgets son bloques que se ordenan en una barra lateral. Cada bloque es independiente del otro. Algunos bloques de ejemplo podrían ser las entradas más recientes, un calendario, una nube de tags, un link a los feeds RSS, etc.

Figura 3. Vista general del WMPU en la administración

1.4 ESTADO ACTUAL DEL WORDPRESS MULTIUSUARIO EN LA UTPL

La plataforma WPMU se implementó en la Universidad Técnica Particular de Loja, en su primera versión, realizada en febrero del 2007, luego se ejecutó una migración de datos a la versión del actual 2.7.1, en junio del 2008, poco a poco el interés de la comunidad universitaria provocó que crearan blogs y llevaran bitácoras de sus actividades académicas y de investigación.

Actualmente existen 1288 blogs distribuidos en blogs de asignaturas, blogs de escuelas de Modalidad Presencial y Modalidad Abierta y a Distancia, blogs de departamentos, blogs de Centros Asociados, blogs personales y blogs de asociaciones.

Cabe mencionar que no existe una categorización y distribución para determinar cuáles blogs pertenecen a las cuatro áreas de la UTPL, ya que carece de políticas para la creación de las direcciones de los blogs en la plataforma del wordpress multiusuario.

1.4.1 Ventajas del WPMU en la UTPL

Las opciones de la implementación del WPMU permiten la descentralización de la administración de cada blog creado bajo el dominio, lo que hace que cada usuario administre y categorice sus publicaciones.

Como ventajas de este proceso se tiene:

- ✓ El administrador puede cargar un número limitado de templates (interfaz) y plugins para la personalización de los blogs por parte de los mismos usuarios, ahorrando tiempo al web máster.
- ✓ Soporta un número ilimitado de blogs y de usuarios
- ✓ Los usuarios registrados pueden interactuar en todos los blogs con la misma clave.
- ✓ Los plugins para todos los blogs de WordPress MU se almacenan en la carpeta plugins en el sitio alojado en el servidor.
- ✓ Permite que el creador del blogs mantenga una gestión integrada, más rápida y sencilla de los mismos.
- ✓ Se puede gestionar varios blogs desde una sola instalación de WordPress, así como la configuración del tipo de archivos que pueden subir en los blogs.

1.5 ESTADÍSTICAS DE ACCESO AL WPMU – UTPL

Para investigar la actividad de los blogs se utilizó la herramienta Google Reader[3] con la finalidad de determinar los niveles de actividad del WPMU-UTPL, se evalúan 230 blogs registrados durante el período desde el 7 de enero al 6 de febrero del presente año, se registran 4 blogs leídos. Referente a los días de la semana que muestra mayor actividad de los blogs, se observa un alza el día jueves, mientras que el fin de semana la actividad disminuye, como se muestra en las imágenes del *Anexo 1*.

1.5.1 Estudio de los blogs de la Plataforma WPMU en la UTPL

Para determinar la muestra del universo de 1288 blogs que serán analizados, se utilizó la fórmula de obtención de muestras para poblaciones finitas menores a 100.000 elementos:

$$n = \frac{Z^2 * P * Q * N}{E^2(N - 1) + Z^2 * P * Q}$$

Formula 1. Obtencion de muestra para estadisticas en blogs

Cuya nomenclatura es:

[3] Google Reader es un lector de RSS y atom. Permite organizar y acceder rápidamente desde una interfaz Web a todas las noticias de las páginas configuradas en el sistema que soporten.

n = Número de elementos de la muestra

N = Número de elementos de la población o universo (Actualmente existen 1288 blogs)

P/Q = Probabilidades con las que se presenta el fenómeno

Z^2 = Valor crítico correspondiente al nivel de confianza elegido; siempre se opera con valor zeta 2, luego Z = 2.

E = Margen de error permitido (determinado por el responsable del estudio)

Porcentaje de certeza del 95%

Margen de error del 5%

$$n = \frac{1{,}96^2 * 50 * 50 * 1288}{5^2(1288 - 1) + 1{,}96^2 * 50 * 50}$$

$$n = 296$$

De la muestra obtenida se analiza n blogs, los cuales representan el 22,98% del total de blogs, y mediante el análisis de observación directa, se revisan: entradas, categorías, etiquetas y actividad de los mismos desde su creación hasta Febrero 2010.

Se tomará en cuenta los siguientes aspectos:

- ✓ Poseer entradas diferentes a la entrada de prueba ejemplo: "Hello World"
- ✓ Se contabilizan las entradas etiquetadas, no etiquetadas, entradas categorizadas, entradas no categorizadas.

Para la evaluación del estado actual de los blogs tienen de las 1174 entradas evaluadas que se encontraron en la muestra obtenida, y en la Tabla 3, se muestra un resumen del análisis.

Tabla 3. Resumen del Estado de Blogs del WPMU-UTPL

Entradas evaluadas	Usuarios	Blogs evaluados	Actividad	
			Activos	**Inactivos**
			12%	88%
			Etiquetación	
			Etiquetadas	**Sin etiquetas**
1174	360	296	38%	62%
			Categorización	
			Categorizados	**Sin Categorías**
			81%	19%
			Archivos	
			Sin archivos	**Con archivos**
			71%	29%

- ✓ Los estados de entrada que posee el WPMU-UTPL se detalla en el Anexo 2, literal 1.
- ✓ Los usuarios que poseen interactividad en los blogs se visualizan en el Anexo 2, literal 2.
- ✓ La actividad de los blogs se puede observar el análisis en el Anexo 2, literal 3.
- ✓ La etiquetación de entradas. Ver anexo 2, literal 4.
- ✓ La categorización de entradas están disponibles en el Anexo 2, literal 5.
- ✓ La cantidad de archivos subidos en las entradas se observan en el Anexo 2, literal 6.

La actividad de los blogs es 12%, siendo mínima con respecto a la cantidad de blogs creados, de los cuales existen el 62% de entradas sin etiquetas, además del 19% sin categorizar.

De los blogs activos se desea extraer la información contenida en las entradas, pero es necesario realizar un análisis de las técnicas de extracción de información que sirvan como base para la realización del sistema de etiquetación que se propone en el presente proyecto.

1.6 EXTRACCIÓN DE INFORMACIÓN DE LOS BLOGS

La importancia de los Sistemas de Extracción de Información (SEI): Los SEI tienen como objetivo obtener información relevante e ignorar la irrelevante de recursos existentes en determinado dominio [3]. Proporcionando accesibilidad al recurso, mejorando la localización, colaboración, integración y recuperación de información de la Web, a través de motores de búsqueda.

La importancia de la etiquetación. Facilita el acceso de Internet a los recursos subidos en el WPMU-UTPL, y según la Figura 1, existe alrededor del 62% de entradas sin etiquetas, formando una caótica organización de la información de recursos alojados en la plataforma.

Cabe destacar que se analiza la extracción de información de los recursos del WPMU-UTPL, determinando si resulta factible o adaptable alguna herramienta para utilizar su código en el desarrollo del sistema de etiquetación automática a las entradas de los blogs.

Se puede amplificar la extracción de información con la aplicación de metadatos, entiéndase como *metadato a datos que describen a otros datos*, ejemplo: de un recurso se pueden extraer metadatos como: *su título, autor, fecha de creación, enlaces, etc.*

Los archivos asociados a los blogs tienen formatos como: Microsoft Office, OpenOffice.org, archivos de Windows 2000, documentos PDF, imágenes. Dentro de los parámetros utilizados para el análisis de los blogs se considera lo siguiente:

- ✓ Entradas: Post ingresados dentro de cada blog
- ✓ Autores: La persona que escribió o publicó un post
- ✓ Categorías: Categorías utilizadas en la ordenación y clasificación de entradas
- ✓ Etiquetas: Tags utilizados para las entradas o archivos
- ✓ Entradas no Categorizadas: Post sin categorizar.
- ✓ Descripción del post: El resumen de que se trata la publicación
- ✓ Documentos: Todo tipo de documentos permitidos en wordpress.

Las herramientas que permiten la extracción de metadatos, tanto para archivos y enlaces de sitios, se pueden representar en formatos representados en XML (Extensible Markup

Language) y RDF (Resource Description Framework), los cuales incluyen atributos y relaciones, que proporcionan un vocabulario consensuado para definir redes semánticas de unidades de información interrelacionadas que especificarán las reglas lógicas para que los agentes de software reconozcan y clasifiquen cada concepto.

1.7 TÉCNICAS DE DEFINICIÒN DE METADATOS COMPATIBLES CON WORDPRESS

La extracción automática de metadatos y la anotación de contenido permiten calificar los contenidos desde las fuentes de datos siendo capaz de extraer metadatos.

Las técnicas de definición de metadatos son:

- ✓ Mediante archivos XML (Ver Anexo 3, literal 1)
- ✓ Mediante archivos RDF (Ver Anexo 3, literal 2)
  - Mediante la aplicación de RDF SIOC (Ver Anexo 3, literal 2.1)
- ✓ RSS (Ver Anexo 3, literal 3)
- ✓ Mediante taxonomías WP TAGS SCHEMA (a nivel de tags) (Ver Anexo 3, literal 4)
  - Aplicación de la Técnica de WP Tags Schema (Ver Anexo 3, literal 4.1)

Las tecnologías que describan de recursos alojados en blogs, se debe obtener como resultado un archivo que recoja dicha información y sirva de alimentación a aplicaciones que representen el conocimiento, se pretende que se utilice un formato compatible e interoperable entre aplicaciones web como es el caso de RDF ó XML, que extraen los elementos necesarios de las entradas de un determinado blog.

1.7.1 Comparación de las tecnologías de extracción de metadatos

Para facilitar la interpretación de la extracción de metadatos, se evalúan las ventajas y desventajas en el uso de las tecnologías, las cuáles se pueden utilizar en la investigación de este tema, como se presenta en la Tabla 4.

Tabla 4. Cuadro comparativo de las tecnologías aplicables al WPMU_UTPL

Aspectos	Mediante XML	Mediante RDF	Mediante RSS	Mediante WP TAGS SCHEMA
Uso de formato estandarizado	Si	Si	Si	No
Totalidad de metadatos consideración en la extracción	Si	Si	No	No
Formato reutilizable en aplicaciones web	Si	Si	Si	No
Datos semantizados	Si	Si	Si	No
Ayuda al acceso de recursos y etiquetación	Si	Si	Si	Si
Uso de estándar de metadatos en la descripción de recursos	No	No	No	No

Una vez evaluadas las técnicas de definición se puede confirmar que debido a los resultados que se presenta se utilizará el RDF, ya que al utilizar un archivo RDF *Resource Description Framework* es una DTD (definición del tipo de documento) de XML, es una aplicación de metadatos que utiliza XML a fin de proporcionar un marco estándar para la interoperabilidad en la descripción de contenidos web.

Según la obtención del resultado del la tecnología SIOC en el Anexo 3, literal 2.1, se ve una extracción de información de manera caótica, donde el RDF no tiene una estructura ordenada, por lo tanto se pretende realizar la investigación de estándares para la definición de metadatos.

Para agilizar el proceso de extracción de metadatos de recursos, a continuación, se realiza una investigación de herramientas que brindan este servicio.

1.8. HERRAMIENTAS DE EXTRACCIÓN DE METADATOS

1.8.1 Herramientas en servidores

Resumen de resultados:

Según la comparación de software entre Foca online y Foca servidor ambas herramientas, realizan la extracion de información. Se deduce que, la versión online tiene mayor funcionalidad, debido a la ventaja en características de mayor número de formatos soportados y el tiempo de respuesta.

El uso de casos de prueba para realizar las comparaciones se obtiene como resultado la siguiente tabla de comparaciones evaluando sus características.

Tabla 5. Tabla comparativa de las características entre FOCA online y FOCA de servidor

Items evaluados	Foca online	Foca servidor local
Datos relativos a las fechas	Si	Si
Tipos de formatos de recursos que son extraídos: .doc .ppt .pps .xls .docx .pptx .ppsx .xlsx .sxw .sxc .sxi .odt .ods .odg .odp .pdf .wpd	+ formatos como .svg .svgz .jpg	Carece de algunos formatos.
Extracción de metadatos de imágenes embebidas en documentos.	Si	No
Tipos de datos que extrae: Título, aplicación, codificación, compañía, numero de ediciones, plantilla, sistema operativo, tiempo de edición, usuarios	Todos	Todos + tamaño de archivo.
Almacenamiento de datos	No almacena ningún dato de los documentos subidos, únicamente valores numéricos de la cantidad de elementos encontrados para mostrar las estadísticas.	Si almacena en el directorio creado para la recuperación.
Tiempo de respuesta	Menor por cada recurso, dependiendo del ancho de banda a utilizarse.	Mayor, debido a que usa una aplicación en servidor local.

Transformación de resumen a archivos RDF.	No	No

Luego de la evaluación de las herramientas en la extración de metadatos de un archivo de word, se resume que, los metadatos devueltos son, como: nombre del autor, tipo de sistema operativo, fechas de actualizaciones y relacionadas con el software. Ver Anexo 4, literal 1.

Debido a que la presentación de resultados obtenidos son metadatos básicos, por lo tanto en este proyecto no aplicaría la reutilización de código.

1.8.2 Herramientas de extracción de metadatos en clientes

En la evaluación de las herramientas de extracción de metadatos en clientes se ha evalauado las características de el software Metadata Miner Catalogue PRO y el software HTML Code Export, y se ha obtenido la siguiente tabla comparativa.

Cuadro comparativo de las características de las herramientas:

Tabla 6. Tabla comparativa de Metadata Miner Catalogue PRO software y HTML Code Export

	Metadata Miner Catalogue PRO software	**HTML Code Export**
Licencia	Si	No
Archivos de Entrada	Archivos de **Microsoft Office** **OpenOffice.org** **HTML** **PDF** imágenes JPEG / PSD	PDF,RTF,BMP,PNG,JPG, Lotus,SVG,QUATTRO Pro,Excel
Archivos de Salida	**HTML** **XML** **RDF** **Informe de exportación CSV** **Informe de MS Word**.	HTML, XHTML and XML, TXT, RTF Diagnóstico, y Codificación de caracteres.
Configuración de salida de metadatos	Si	Si
Modificación de la propiedad de los archivos	Si	Si

Según la comparación de la Tabla anterior se deduce que debido a que el Software Metadata Miner Catalogue Pro es de versión propietaria, y no se tiene presupuestado la adquisición de su licencia; comparado con las características y las limitantes del Software HTML Code Export cuyos archivos de salida no existe el formato RDF o XML. Ver Anexo 4, literal 2.

A continuación se procede a determinar que:

- ✓ El número de archivos alojados en el WPMU-UTPL es minoritario, y se procede a la evaluación de contenido de las entradas sin etiquetar en cuyo interior se encuentran archivos.
- ✓ Al evaluar archivos y extraer su dirección dentro de una entrada de determinado blog se redundaría en la obtención de la URL, debido a que si se devuelve solo el enlace del archivo, lo que limita la visualización del contenido presente en la entrada donde se aloja dicho archivo.

Para la extracción de información de contenido de las entradas del WPMU-UTPL se considera las entradas que carecen de etiquetas, realizándose un análisis de las herramientas que faciliten dicha extracción y organice la información en un formato que se pueda describir en archivos semánticos y enriquezca la estructuración de la información para ser procesados de forma automática.

Existen herramientas o que ayudan en la evaluación de contenido y devuelven las palabras relacionadas o palabras clave o tags.

1.9 HERRAMIENTAS DE EXTRACIÓN DE ETIQUETAS ONLINE

Según el análisis de las entradas sin etiquetas en los blogs del WPMU-UTPL es el 62%, debido la carencia de tags o palabras clave que el usuario le brinda a sus publicaciones.

En la extracción automáticamente de información estructurada o semiestructurada desde sitios web, actualmente existen técnicas de procesamiento de lenguaje natural.

Siendo el Reconocimiento de Nombres de Entidades (NER[4]), conocida como la identificación de entidad y de extracción de la entidad, es una subtarea de extracción de información que busca localizar y clasificar los elementos atómicos en el texto.

Los sistemas NER utilizan técnicas basadas en la gramática lingüística, y modelos estadísticos; que haciéndolos de forma manual, costaría de meses de trabajo, gasto de recursos, para los experimentos lingüísticos computacionales.

Se puede recurrir a un servicio web que ayude a la extracción de tags a partir de la evaluación de texto mediante el uso de APIs disponibles y configurables para implementarse en desarrollo o adaptación de aplicaciones que tengan por objetivo la recuperación de información y de cómo resultado en un formato entendible para la adaptación al sistema a desarrollar.

Existen herramientas en línea que permiten la extracción de etiquetas como:

- ✓ Open Calais
- ✓ AlchemyAPI

Estas son herramientas que permiten realizar la extracción de tags a partir de la evaluación de contenido de una URL. Ver Anexo 5, literal 1 y 2.

A continuación se presenta la evaluación de la comparación de ambas herramientas.

Comparación de los extractores de etiquetas

[4] Una técnica es el reconocimiento de nombres de entidades, que utilizan técnicas basadas en la gramática lingüística y modelos estadísticos.

Tabla 7. Comparación de herramientas de extracción de etiquetas online

Items Evaluados	Open Calais	AlchemyAPI
Reconocimiento de Entidad (NER)	Si	Si
Limitación en el tamaño del texto evaluado	Si	No
Utilización en varios idiomas	Si	Si
Usa PLN Procesamiento de Lenguaje Natural, aprendizaje de máquinas y otros métodos	Si	Si
Extrae metadatos semánticos	Si	Si
Incorporación con otras aplicaciones	Si	Si
Key para usar el API	Si	Si
Distribución gratuita	Si	Si
Extracción semántica de datos	Si	Si
Respuestas de Formatos	RDF Microformatos Formato simple JSON	RDF XML JSON
Transacciones por día	50.000 transacciones por día	30.000 llamadas a la API cada día.
Métodos admitidos por los Web Service	SOAP REST	SOAP REST

Según la evaluación y comparación de las características de ambas herramientas, se deduce que, es más aplicable el Web Service Alchemy API, por la nitidez de los tags devueltos.

1.10 SOLUCIÓN EN LA ADQUISICIÓN DE ETIQUETAS

Según la investigación, se determina que, para las entradas que no poseen etiquetas en blogs, es posible hacer uso de una herramienta online Alchemy API que permite evaluar el texto, y mediante una API es posible la implementación y adaptación a la aplicación solución que se brindará en este proyecto.

Se conoce que Alchemy API realiza un análisis de contenido de texto en sitios web, utiliza el reconocimiento de entidad llamado (NER) además de algoritmos estadísticos y lingüísticos avanzados que analizan el contenido, "etiquetándolo" con las palabras más importantes. Además de ser un web service, se puede usar en otras aplicaciones mediante sus SDK[5] disponibles en su sitio oficial en diferentes lenguajes de programación.

[5] Un kit de desarrollo de software, que le permite a un programador crear aplicaciones para un sistema concreto.

Para utilizar el servicio online se debe obtener una clave API, y a partir de ésta adpatación se interpreta sus parámetros de envio (link de entrada de blog) y recepción (tags).

Una vez obtenidos los datos necesarios para la descripción de las entradas del WPMU-UTPL (bookmarks[6]), con los campos que lo identifiquen, y a su vez sea un archivo semántico el que se utilice como producto final de la extracción de información de los contenidos de blogs.

Marcadores sociales como: Del.icio.us[7], Digg[8], Stumbleupon[9] y Reddit[10], facilitan la difusión de información y permite que personas con intereses similares puedan ver los enlaces por categorías, etiquetas o al azar de temas de su interés.

[6] Un bookmark en la World Wide Web, es una dirección Web dirección WWW o URL que queda archivada para su posterior uso, para marcar una Web interesante a fin de poder volver a él posteriormente. (ver Delicious)
[7]Marcador Social Delicious: http://www.delicious.com/
[8] Marcador Social Digg: http://digg.com/news
[9] Marcador Social Stumbleupon: http://www.stumbleupon.com/
[10] Marcador Social Reddit: http://www.reddit.com/

CAPITULO 2

2. IMPLEMENTACIÓN Y ADAPTACIÓN DEL SISTEMA EMEB

INTRODUCCIÓN

En el presente capítulo se describe en detalle la estrategia seleccionada como solución de la integración del sistema de marcadores a partir de la extracción de información de entradas de los blogs de la UTPL. Esta adaptación contempla la incorporación de tablas en la estructura de la base de datos del WPMU-UTPL.

OBJETIVOS:

General:

- ✓ Desarrollar un sistema semiautomático de etiquetación de contenidos de blogs utilizando lenguajes de descripción semántica.

Específicos:

- ✓ Lograr la extracción de metadatos de las entradas de blogs.
- ✓ Etiquetar las entradas sin etiquetas para el proceso de extracción.
- ✓ Generar bookmarks con el resultado de la extracción de metadatos de los blogs.

RESULTADOS ESPERADOS:

Los resultados a obtener en esta fase

1. Sistema de obtención de metadatos.
2. Reutilización de recursos subidos en el wordpress MU de la UTPL, en plataformas de bookmarks.
3. Archivos de bookmarks en formato RDF, obtenidos de los post publicados en el wordpress MU.

ESTRATEGIA:

La estrategia que se propone para este proyecto comprende los siguientes pasos:

- ✓ Desarrollar la programación en PHP para la extracción de etiquetas de la base de datos del WPMU y la generación de archivos RDF.
- ✓ Adecuar una taxonomía en base a la distribución de areas de la universidad, y que ayudará a la organización de la información.
- ✓ Utilizar el web service AlchemyAPI como extractor de metadatos para las entradas sin etiquetas.
- ✓ Generar automáticamente un archivo RDF que estructure los bookmarks obtenidos de la extracción de los blogs.
- ✓ Realizar pruebas de funcionamiento del proceso de extracción de metadatos.
- ✓ Adaptar la aplicación en un servidor de pruebas con la finalidad de prevenir algún error posterior a la implementación.

BENEFICIOS:

- ✓ Capacidad de reutilización de la información extraída desde los blogs del WPMU-UTPL en cualquier sistema de representación de conocimiento.
- ✓ Organización de la información de los blogs de acuerdo a su contenido.
- ✓ Obtención de archivos RDF, que puede ser registrado en buscadores semánticos y utilizados por herramientas de representación semántica.

2.1 ESTRATEGIA DE INTEGRACIÓN

2.1.1 Requerimientos del Sistema EMEB

Es primordial que se extraiga los metadatos que describen a las entradas de los blogs y luego generarlos en bookmarks mediante archivos RDF.

Luego del análisis de la situación actual de los blogs se determina que existen entradas cuyos metadatos son incompletos como es el caso de las etiquetas parte fundamental en este proyecto; en consecuencia, se definen los requerimientos para la gestión apropiada de del Sistema EMEB[11]. Igualmente se necesita que sus procesos sean de ejecución automática, mediante rutinas KSH[12].

En cuanto a la especificación de requerimientos se encuentran los siguientes:

- ✓ REQ 001. Extracción de metadatos, cuando la entrada posea tags.
- ✓ REQ 002 Extracción de metadatos, cuando la entrada no posea tags.
- ✓ REQ 003 Generación de Archivos RDF

Los estados de los requerimientos como: la entrada, proceso y salida se describen en el Anexo 6.

2.1.2 Definición de el Problema

El problema de:	Según los procesos de estudio y análisis, en los blogs alojados en el WPMU-UTPL se deduce que se conserva excelente información que aporta al conocimiento a usuarios que utilizan plataformas virtuales como apoyo para su estudio, existe gran cantidad de blogs que son escasamente etiquetados y, actualmente no existe una categorización a nivel superior, que permita realizar filtros temáticos.
Afecta a:	El sistema tiene la característica de ser oculto, el mismo que solo es percibido por el administrador de la plataforma, pero se ven involucrados todos los usuarios que ingresen, modifiquen, borren entradas dentro de cualquier blog; a continuación se detallan los usuarios que se involucran en modo oculto con el Sistema EMEB: ✓ Administrador del wordpress multiusuario (directamente) Usuarios que ingresan post (entradas) de wordpress. (indirectamente)
Cuyo impacto es:	Con el aporte de etiquetas a las entradas de blogs, mejora la accesibilidad a la información que se esconde en cada uno de los blogs.
Una solución exitosa es:	Adaptar un Sistema de bookmarks automático, tomando como base a las entradas ingresadas en el blog multiusuario de la UTPL, que permita la descripción semántica con información relevante de su contenido y provea informaciòn hacia aplicaciones de

[11] Sistema de Etiquetación de Metadatos a Entradas de Blogs
[12] El Korn shell (ksh) es un programa informático cuya función consiste en interpretar órdenes por líneas.

	representación de conocimiento.

2.1.3 Posicionamiento del producto:

Para	Administradores web, además de involucrase indirectamente los usuarios que aporten con entradas al WPMU-UTPL
Quien(es)	Desean consumir el material educativo alojado en los blogs del multiusuario de la UTPL. Usuarios que requieran categorizar por temática las publicaciones realizadas en el multiusuario. Administradores de aplicaciones de representación de conocimiento a través de bookmarks que deseen presentar hacia cualquier sitio web la información de las entradas de los blogs desde la UTPL.
El Nombre del Producto	Dentro de la plataforma wordpress se adaptará el Sistema de Etiquetación de Metadatos a Entradas de Blogs (EMEB)
A Diferencia	De la forma actual como se encuentran los recursos educativos alojados en el wordpress multiusuario se encuentra la información sin categorizar, permaneciendo no accesibles por la mayoría de usuarios, el sistema propuesto debe ser: Lo más natural posible de manera que no altere su acostumbrada interacción con el blog que ingrese nuevas entradas. Oculto al usuario final, de manera que no tenga inconveniente en su interacción con los blogs que estaba usando. Para la creación de nuevos blogs se propone la adaptación de un campo "Combo Box" para la creación de nuevos blogs, para que categoricen de acuerdo a la temática que tenga un blog. Procesos automáticos de manera que no dependa de un administrador para la ejecución de algún proceso interno propio del Sistema EMEB
Esta Aplicación:	Proporciona la capacidad de: ✓ Permitir a la comunidad universitaria mantener la descripción semántica de recursos desde el wordpress multiusuario en archivos RDF. ✓ Permitir al administrador mantenerse informado de los bookmarks extraídos, y especificaciones de la cantidad de bookmarks generados a archivos RDF. ✓ Permitir al administrador mantenerse informado de los tags extraídos a las entradas que no poseen etiquetas. Servir de fuente de información de educación para integrarla en sistemas de representación de bookmarks.

2.1.4 Casos de Uso del Sistema EMEB

El proceso general se basa en una serie de pasos cuyo transcurso es el siguiente:

- ✓ CU001- Selección y Adquisición de Metadatos
- ✓ CU002- Extracción o generación de etiquetas (Consumo del web Service a través del API)
- ✓ CU003- Almacenamiento en la Base de Datos
- ✓ CU004- Generación de RDF

El curso de los casos de uso es detallado en el Anexo 7.

2.1.5 Rol de Usuarios del Sistema EMEB

El sistema EMEB posee solo un rol de usuario, para el resto de usuarios del wordpress será imperceptible.

- ✓ Grupo Administradores

Todos los usuarios que estén dentro del grupo de "Administrador" podrán realizar cualquier actividad dentro del wordpress multiusuario, como: (lectura, escritura, edición, administración) sobre los bookmark obtenidos.

En la implementación, se determinó que únicamente el grupo del rol "Administrador" sean quienes acuden a la información del resultado del Sistema EMEB.

2.2 CONSIDERACIONES CON LOS DATOS DE LAS ENTRADAS DE LOS BLOGS ACTUALES

Para la extracción de metadatos no se ha tomado en consideración el blog 1 y el post 1 que corresponde al blog multiusuario y al "Hola Mundo" respectivamente.

Dado a que las entradas actuales no comprenden todos los datos necesarios, lo que hace que dos campos de descripción de bookmarks se presente de forma incompleta, los cuales son:

- ✓ Entradas sin tags asociados.
- ✓ No existencia de categorías de nivel superior.

2.2.1 Entradas sin tags asociados

Para el llenado de tags en caso de que la entrada no posea se usa el web Service Alchemy API que es uno de los más interesantes, para el cual se usa el "SDK Alchemy API que utiliza PHP Versión 5.0" que se adapta a sistemas, donde sea necesario un análisis de lenguaje natural y extracción de palabras clave de un determinado contenido y la adaptación al Sistema EMEB.

2.2.2 Adaptación e implementación del Web Service Alchemy API para generar etiquetas

Alchemy API utiliza la tecnología de procesamiento del lenguaje natural y algoritmos de aprendizaje automático para analizar contenido en páginas web accesibles desde Internet, publicado el contenido en HTML o texto, e imágenes de documentos escaneados, y realiza la extracción semántica de tags o palabras clave, ya que utiliza algoritmos estadísticos avanzados y lingüística para el análisis de su contenido de palabras más importantes.

Por lo general no existe la cultura por parte del usuario para agregar etiquetas a sus publicaciones, es importante que se tome en cuenta este tipo de casos, ya que para el apropiado uso del Sistema EMEB, se debe cumplir con la extracción de todos los datos necesarios para la descripción de bookmarks.

Por tal razón se pretende hacer uso del Web Service AlchemyAPI, el mismo que realiza el proceso de generación de etiquetas a partir de las entradas del blog, devolviendo los tags en el formato XML, el mismo que puede ser utilizado en programas que tengan afinidad con los tags.

Según el ámbito de uso de este Web Service se utiliza como un sustento semántico para la generación de etiquetas a partir del análisis de una entrada de blogs específica, cuyo resultado de tags generados sirve como un elemento importante en la utilización de tags que son almacenadas en el campo bTags de la tabla wp_bookmarks, que almacena la información necesaria para la generación de bookmarks.

Se necesita una clave de acceso, que se la obtiene en el sitio oficial del Web Service. Para utilizar AlchemyAPI se debe realizar el estudio de las características y pruebas.

- ✓ Para las llamadas al API. *Ver Anexo 8, literal 1.*
- ✓ Los parámetros de AlchemyAPI. *Ver Anexo 8, literal 2.*
- ✓ Los formatos de Respuesta de AlchemyAPI. *Ver Anexo 8, literal 3.*
- ✓ La respuesta de campos de AlchemyAPI. *Ver Anexo 8, literal 4.*
- ✓ La implementación y prueba de Alchemy API para la extracción de Tags. *Ver Anexo 8, literal 5.*
- ✓ Extracción de tags mediante AlchemyAPI para el Sistema EMEB. *Ver Anexo 8, literal 6.*
- ✓ Características a considerar del API. *Ver Anexo 8, literal 7.*
- ✓ Los idiomas soportados se pueden en el *Anexo 8, literal 8.*

2.2.3 No existencia de Categorías de Nivel Superior

Para la obtención de las categorías, cabe mencionar que actualmente el wordpress multiusuario no posee una organización a nivel general de los blogs y además no existe un formato para los nombres de los blogs. Ejemplo:

- ✓ http://blogs.utpl.edu.ec/seguridadindustrialeimpactoambientaldeactividadesdelaiet
- ✓ http://blogs.utpl.edu.ec/iaavanzado/

Debido a que no existe una norma para la creación de nombres de los 1.288 blogs alojados hasta el momento en el wordpress multiusuario de la UTPL, y no está estipulada la realización de un sistema que clasifique dichos blogs, según el nombre.

La organización temática de los blogs se realiza manualmente de aquellos que estén activos, tornándose tediosa y lenta la clasificación, pero que una vez categorizados permite obtener la información ordenada para la posterior representación por sistemas de bookmarks.

Para este proyecto se realiza la adaptación del plugin Category Mapping que permite organizar los blogs a varios niveles de categorías.

Cabe mencionar que, con la implementación de plugin Category Mapping, se adapta una tabla adicional en todo el WMPU_UTPL, que contiene la información de los siguientes campos como se muestra en la Figura 4.

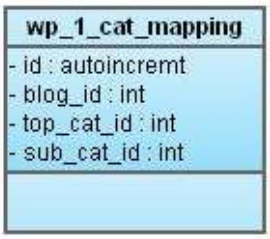


Figura 4. Tabla creada por el Plugin Category Mapping

Luego del proceso de implementación del Sistema EMEB al wordpress multiusuario, cada vez que un usuario cree un nuevo blog se añadiría un botón para la categorización del blog. A continuación se explica la propuesta para su clasificación.

2.2.3.1 Propuesta para categorizar de los blogs en el multiusuario

Actualmente no existe la categorización de los blogs, es necesario que se organice la información de que tipo de recursos corresponden a la aplicación de contexto y que distinga del tipo de blogs están dentro de temáticas de conocimiento.

Según el análisis de los blogs su organización es caótica, debido a esta razón, se propone una taxonomía explicada en la siguiente sección, que sirve para la categorización según el ámbito del tema del blog.

Para la aplicación de categorías que permitan tener un nivel superior, se implementa una extensión o también llamado plugin Category Mapping[13], el mismo que hace uso de las subcategorías ya adjuntas, y se categorice a nivel superior lo que resulta útil para realizar filtrados por categorías o por temas acordes a la información de los blogs.

2.2.3.2 Taxonomía para la clasificación de categorías de los blogs

La aplicación de taxonomía consiste en crear un glosario de términos que pertenecen al dominio (estableciendo una clasificación o jerarquía entre los conceptos, sus niveles, las relaciones entre ellos, sus instancias, sus propiedades o atributos, e igualmente los axiomas o reglas).

Según el ámbito del contenido de los blogs del WPMU-UTPL, se plantea organizar los blogs de acuerdo a las 4 áreas académicas como: Área Socio-Humanística, Técnica, Biológica y Administrativa; además de la existencia de varios de blogs de Centros Asociados, Blogs de Departamentos de la UTLP, Blogs Personales, y blogs de Asociaciones como: grupos estudiantiles, grupos de profesores o de intereses similares.

[13] Es un widget o extensión de código del wordpress

A continuación se presenta la taxonomía que se implementa para la categorización de los blogs, lo que permite organizar la información para obtener información catalogada según intereses de filtración de bookmarks.

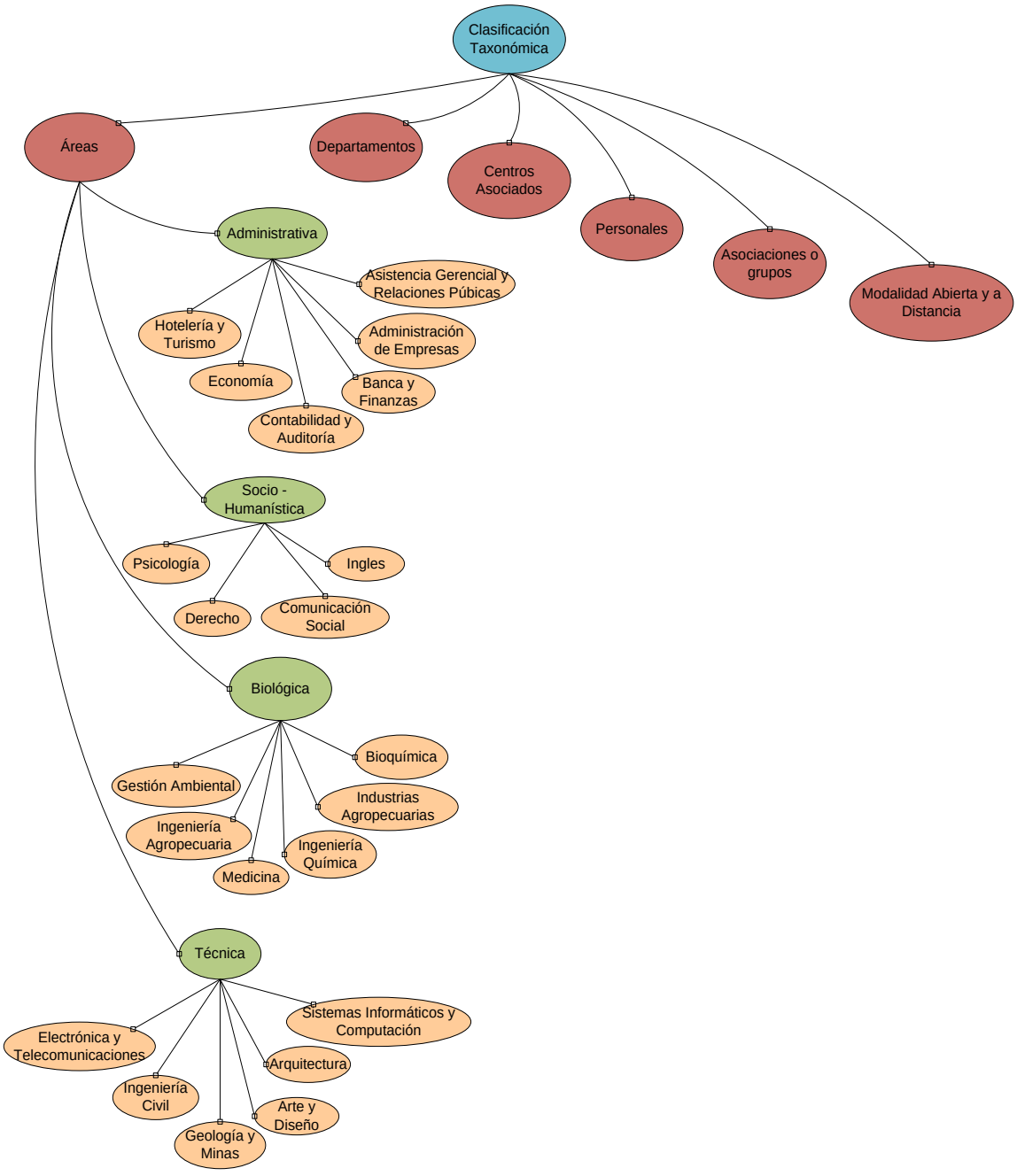


Figura 5. Taxonomía base para la categorización de blogs del WPMU UTPL

El proceso de instalación, configuración y prueba del plugin Category Mapping. *Ver Anexo 9.*

Además de la clasificación de los blogs con la taxonomía anterior, es necesario crear un campo adicional en el formulario de creación de nuevos blogs para que el usuario escoja la categorización, así como mantener organizada la información facilitando el uso de filtros

necesarios por temática, por área, por materia, por centros asociados, por departamentos, y grupos afines. Ejemplo:

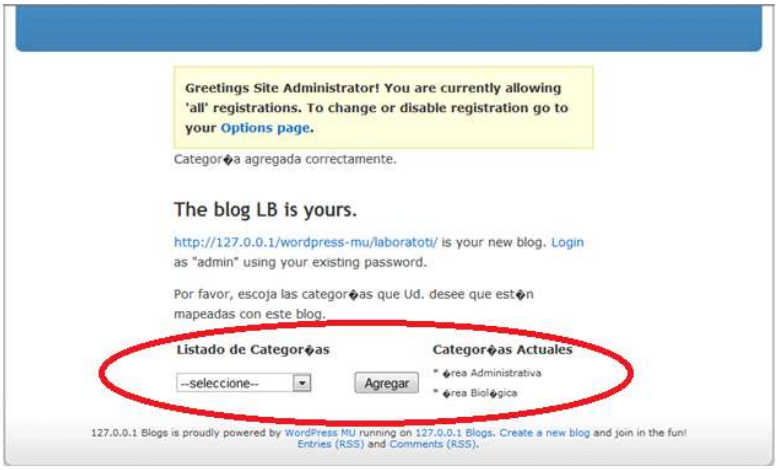


Figura 6. Añadir un campo al momento de crear un blog nuevo, para que el usuario lo categorice

Se realiza cambios en el formulario *wp-signup.php* del wordpress para que el usuario catalogue a su nuevo blog. Ver Anexo 10.

2.3 ESTANDARIZACIÓN DE METADATOS

La falta de conocimiento en aspectos semanticos por parte de desarrolladores de proyectos semánticos, al no aplicar normas a modelos de descripción y definición de metadatos en recursos y la debilidad de la indización de la información digital ocaciona que los recursos que aportan al conocimiento sean invisibles.

Desde la época medieval utilizaban sistemas no estandarizados para mejorar su ordenamiento y búsqueda como la catalogación de bibliotecas, los mismos que solo eran entendidos por el personal de la biblioteca. En vista de aquella situación caótica de desorganización se empezó a desarrollar la normalización que identifiquen los recursos a partir del año 1960,[4] permitiendo intercambiar y compartir el acceso a proyectos cooperativos.

En este contexto, cuando se habla de metadatos, y sobre la base de XML, se definen en distintos lenguajes de marca para los diferentes tipos de documentos; uno de estos lenguajes de marcado semántico es el Resource Description Framework (RDF) como una alternativa para la descripción o "catalogación" de recursos web y por ende, como un modelo de metadatos para mejorar la recuperación de información.

- ✓ El ciclo de vida de los metadatos comprende la creación, gestión propagación y uso. Ver Anexo 11, literal 1.
- ✓ En Internet se distinguen los tipos de estándares, están detallados en el Anexo 11, literal 2.
- ✓ En cuanto a los estándares de metadatos más utilizados por la comunidad esta en el Anexo 11, literal 3.
- ✓ La clasificación de los modelos de metadatos, están expuestos en el Anexo 11, literal 4.
- ✓ La normas o estándares para el uso de metadatos, se describe en el Anexo 11, literal 5.

2.3.1 Estándar a utilizar en la Definición de Metadatos

Para el desarrollo del Sistema EMEB, se utilizará el método descriptivo, pues se busca describir los metadatos más importantes de las entradas de los blogs, y complementado con el método estructural que intervienen en la recuperación de información electrónica como es el caso del RDF[14].

El uso de metadatos debe responder a un estándar utilizado para la definición de cualquier recurso de la forma que sea entendible y escalable entre sistemas afines [5].

2.3.1.1 RDF

El lenguaje RDF es la infraestructura para la descripción de recursos utilizado en situaciones en las que la información necesita ser procesada por aplicaciones que intercambian información, [6] además proporciona una infraestructura que soporta actividades de metadatos.

RDF también provee una sintaxis basada en XML, llamada RDF/XML, para guardar e intercambiar la información.

Es el indicado para definir metadatos sobre recursos web, tales como el título, autor, ect., y ampliamente aceptado por sus características como: idependiente, intercambiable y escalable.[7]

RDF está basado en la idea de que los objetos a describir poseen propiedades que a su vez tienen valores. Estos objetos pueden ser descritos formulando "declaraciones" que especifican estas propiedades y valores y por consiguinete en expresiones con la forma sujeto-predicado-objeto conocidas como tripletas.

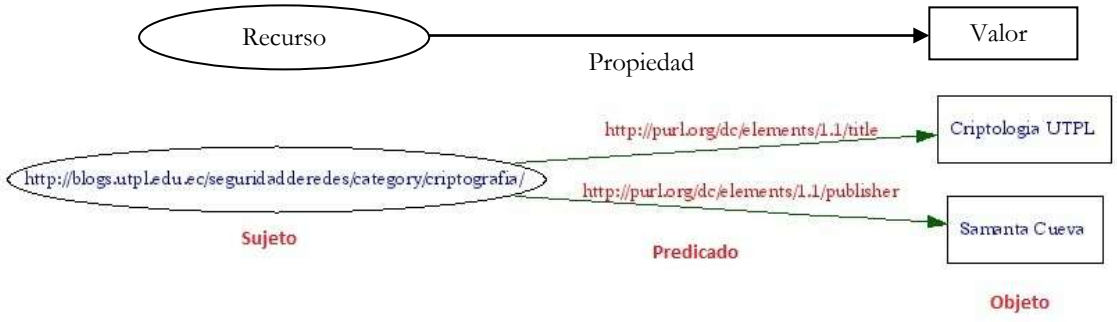


Figura 7. Tripleta de un recurso

Estructura de un documento RDF

RDF toma un tanto de terminología abstracta y otro tanto de sintaxis XML para definir los documentos, de manera que se puedan escribir programas para procesarlos. [8]

Ejemplo:

```
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
```

[14] Resource Description Framework, desarrollado por el W3C para la descripción de recursos web.

```
xmlns:dc="http://purl.org/dc/elements/1.1/">
<rdf:Descriptionrdf:about="http://es.wikipedia.org/wiki/Tim_Berners-Lee">
<dc:title> Tim Berners Lee </dc:title>
<dc:publisher>Wikipedia</dc:publisher>
</rdf:Description>
</rdf:RDF>
```

En este ejemplo se define el recurso cuyo titulo es Tim Berners Lee y que ha sido publicado por la Wikipedia.

2.3.2 Dublin Core

Para la descripción de objetos de información , la DCMI (Dublin Core Metadata Initiative), se define por su norma ISO 15836 del año 2003, que delimita el Conjunto de Elementos Dublín Core [9], dedicada a la participación extensa de estándares interoperables de los metadatos.

Además que es el estándar internacional más conocido para la descripción de los recursos, como elementos básicos para describir cualquier objeto de información.

Las implementaciones de Dublín Core usan generalmente XML.

Dentro de sus ventajas se menciona a continuación: [10]

- ✓ La simplicidad
- ✓ La flexibilidad
- ✓ La independencia sintáctica
- ✓ La interoperabilidad semántica
- ✓ Alto nivel de normalización formal
- ✓ Define el marco para la interoperabilidad entre conjuntos de metadatos.
- ✓ Consenso internacional
- ✓ Modularidad de Metadatos en la Web.
- ✓ Facilita el desarrollo de conjuntos de metadatos específicos de una disciplina o comunidad que trabaja dentro del marco de la recuperación de información

Su objetivo es elaborar normas interoperables sobre metadatos y desarrollar vocabularios especializados en metadatos para la descripción de recursos que permitan sistemas de recuperación más inteligentes.

El conjunto de elementos de metadatos Dublín Core es un conjunto utilizado para describir documentos. Los elementos poseen etiquetas descriptivas que pretenden transmitir un significado semántico a los mismos, además de que estos elementos se pueden repetir, y aplicar en cualquier orden.

Podemos clasificar el conjunto de elementos Dublin Core en 3 grupos que indican la clase o el ámbito de la información que contienen:

- ✓ Elementos relacionados principalmente con el contenido del recurso:
  - **Title** (titulo)
  - **Subject** (tema)
  - **Description** (descripción)

- **Source** (fuente)
- **Languaje** (lenguaje)
- **Relation** (relación)
- **Coverage** (cobertura).

✓ Elementos relacionados principalmente con el recurso cuando es visto como una propiedad intelectual:

- **Creator** (autor)
- **Publisher** (editor) y, otras colaboraciones
- **Contributor** (otros autores/colaboradores)
- **Rights** (derechos).

✓ Elementos relacionados principalmente con la instanciación del recurso:

- **Date** (fecha)
- **Type** (tipo de recurso)
- **Format** (formato)
- **Identifier** (identificador)

Existen proyectos en organizaciones vinculadas con la educación c omo universidades, organizaciones y bibliotecas que actualmente usan la especificación para la descripción de sus recursos con Dublín Core. *Ver Anexo 12.*

2.3.3 Elementos de Dublín Core a utilizar para la descripción de recursos en el sistema EMEB

Al momento de extracción de información mediante la conexión y las consultas a la base de datos del WPMU se debe obtener cono producto la conexión directa con SCUTTLE con los siguientes campos:

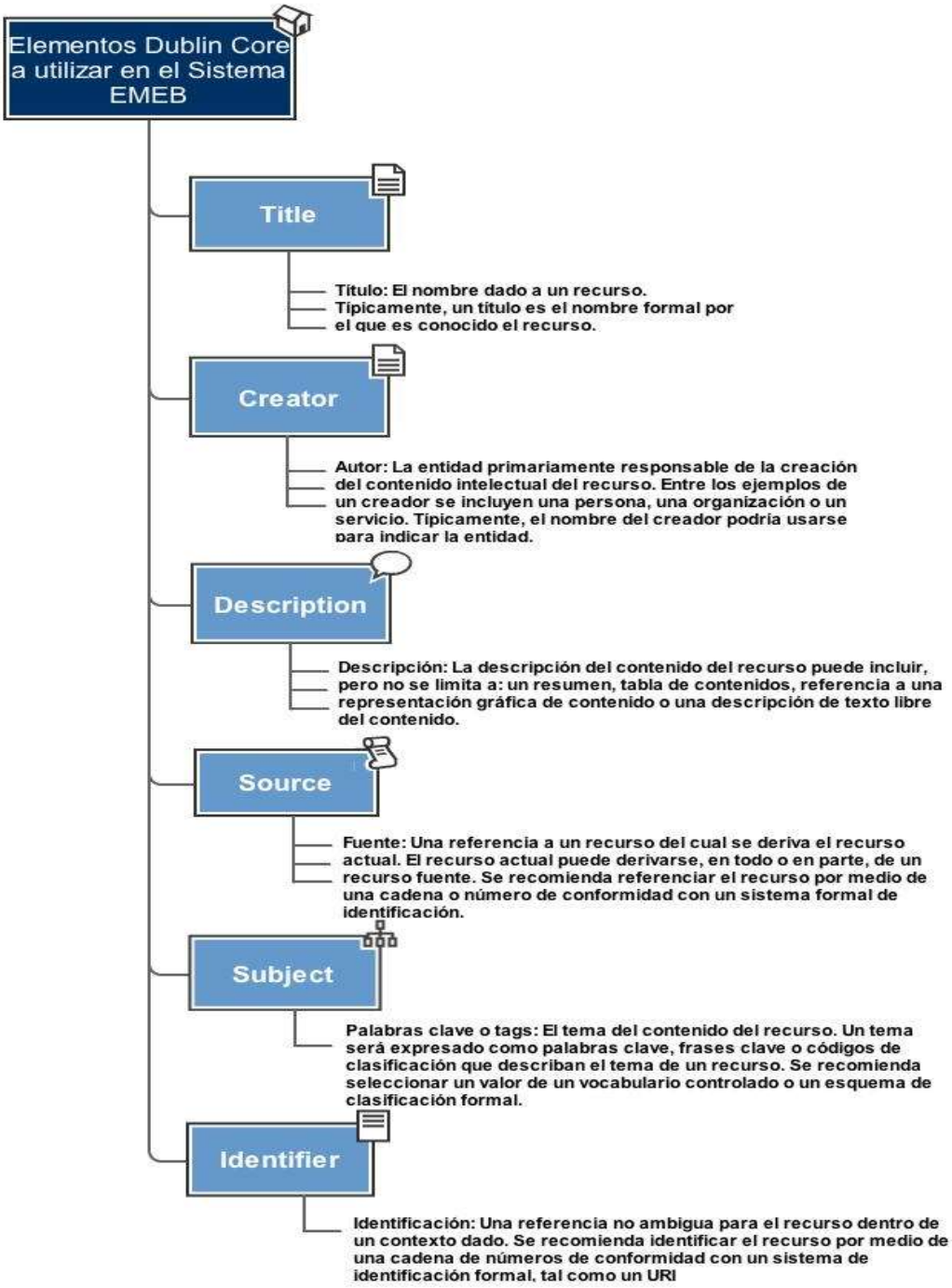


Figura 8. Elementos de Dublín Core propuestos para la descripción de recursos en el Sistema EMEB

Para la implementación del sistema de extracción de metadatos del presente proyecto se usan las especificaciones de Dublín Core.

Los elementos de Dublin Core a utilizar en el Sistema EMEB son: **title, creator, description, source, subject, identifier.**

De cada una de las entradas que se encuentran alojados en los blogs del wordpress multiusuario, la siguiente pantalla representa un blog con los elementos para la descripción de bookmarks.

Figura 9. Una entrada de un blog del multiusuario de la UTPL

Equivalencia de elementos con los metadatos de los recursos del wordpress

Tabla 8. Especificación de los elementos del DublinCore para la descripción bookmarks a utilizar de una publicación de wordpress.

Elementos DC	Elementos del wordpres	Ejemplo:
DC:title	Título de la Entrada de cada blog.	*Instituto de Pedagogía para la Educación y a Distancia*
DC:creator	El autor que realizó la creación de una entrada.	*ipedadmin*
DC:description	Los primeros caracteres de la publicación.	*El 21 de julio de 2010, la Universidad Técnica Particular de Loja (UTPL) y GRANA-ISTEC™ puso en marcha el proceso de evaluación para el programa de UTPL en ingeniería eléctrica e informática.*
DC:source	Es el link o nombre de la publicación (marcador).	http://blogs.utpl.edu.ec/iped/2010/08/17/grana-lanzamineto-en-utpl/
DC:subject	Se utiliza los tags que hayan sido ingresados por los autores de las entradas.	*distancia, educación, iped, utpl*
DC:identifier	Se escogen las categorías ingresadas por el usuario en dicha publicación.	*Maestría EaD, Noticias*

2.3.4 Esquema de un Bookmark

La estructura del bookmark que describe a las entradas de cada uno de los blogs del WPMU-UTPL tiene la siguiente estructura.

Figura 10. Estructura de un Bookmark

2.4 MODELO PROPUESTO PARA EL WPMU-UTPL

2.4.1 Integración de Wordpress Multiusuario con Scuttle[15]

Actualmente se tienen dos sistemas por separado:

- ✓ **WORDPRESS:** es un sistema de gestión de contenido enfocado a la creación de blogs (sitios web periódicamente actualizados).
- ✓ **SCUTTLE**: Es un servicio de gestión de marcadores sociales en web. Permite agregar los marcadores que clásicamente se guardaban en los navegadores y categorizarlos con un sistema de etiquetado denominado folcksonomías (tags). No sólo puede almacenar sitios webs, sino que también permite compartirlos con otros usuarios.

ESQUEMA ACTUAL

Figura 11. Esquema actual entre WPMU-UTPL y Scuttle

Para la integración de los sistemas se debe realizar la adaptación de dos tablas al WPMU, que almacenará la información necesaria para la generación de archivos RDF.

[15] Sistema de Representación de bookmarks

Las principales razones por las cuales se ha escogido estos dos sistemas son:

- ✓ Utilizan el mismo tipo de base de datos (Mysql)
- ✓ Son desarrollados bajo el mismo lenguaje de programación (PHP).
- ✓ Ambos son gratuitos, no necesitamos pagar ninguna licencia para utilizarlos o modificarlos de acuerdo a nuestra conveniencia.

El sistema que se adaptará al WPMU-UTPL se denomina **EMEB** que significa Sistema de Etiquetación de Metadatos a Entradas de Blogs.

2.4.2 Esquema Propuesto para la Implementación e Integración entre EMEB y Scuttle

En el esquema propuesto se realizará la integración de ambos sistemas de la siguiente manera:

Figura 12. Esquema propuesto para la implementación e integración de Scuttle y EMEB

El Sistema EMEB se ubica dentro del wordpress multiusuario en donde se almacenará sus metadatos extraídos, siguiendo con la generación de un archivo semántico RDF. Esta información servirá de suministro para el sistema de representación de bookmarks Scuttle desarrollado en otro trabajo de fin de carrera.

2.4.3 Implementación del Sistema EMEB

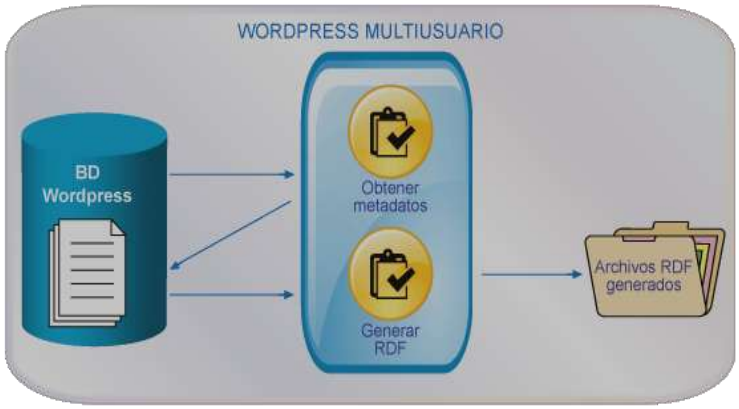


Figura 13. Esquema del Sistema EMEB

Como podemos observar en la Figura 13, el Sistema WPMU permanecerá intacto, no se realizará ninguna modificación a las tablas, sino más bien se añadirán las tablas necesarias para la extracción de datos.

Así mismo el desarrollo será transparente para el usuario del wordpress, no se presentará ninguna diferencia ya que la programación y extracción de datos no implica interfaces para el usuario final.

La programación será desarrollada en PHP, y se utilizará la misma base de datos mysql con pequeñas adaptaciones, de esta forma facilitará el desarrollo al tener los mismos lenguajes de programación y la adaptación a la base de datos.

En cuanto a las entradas que no posean tags se utilizará el web service Alchemy Api para la generación de etiquetas. A partir de la evaluación del contenido de las entradas, de esta manera tendremos los datos depurados, consistentes y listos, para utilizarlos como parte importante en la descripción de bookmarks.

2.4.4 Arquitectura del Sistema EMEB - UTPL

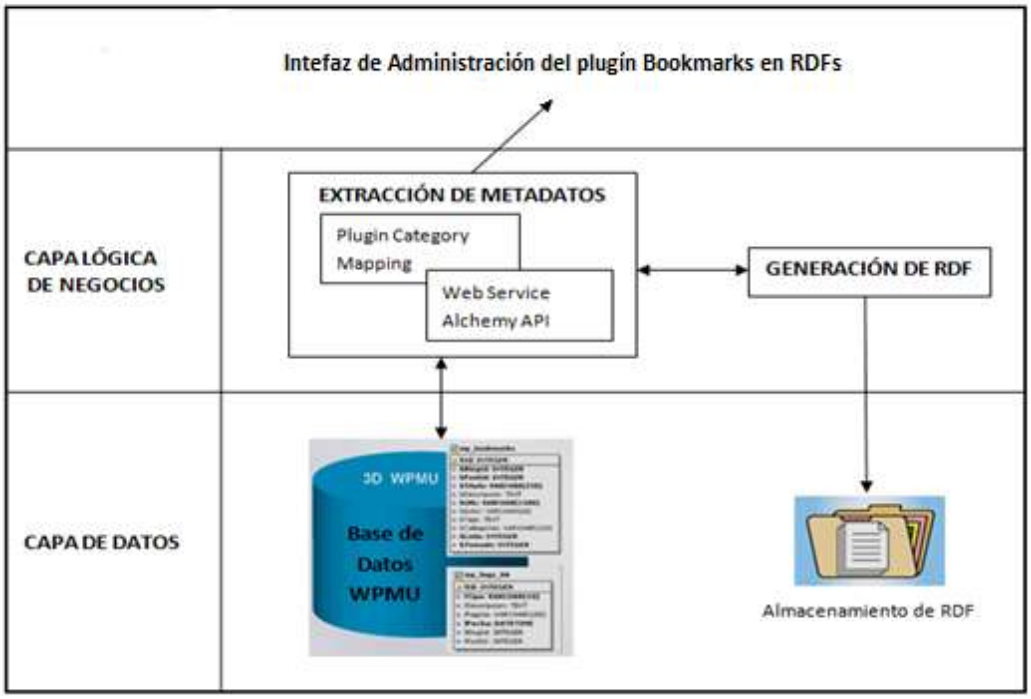


Figura 14. Arquitectura del Sistema EMEB

La arquitectura propuesta del Sistema EMEB, tiene como propósito identificar las partes que intervienen en el desarrollo, y que darán un resultado exitoso. Se ha dividido en tres partes importantes las mismas que se describen a continuación:

Capa de Presentación

Aunque el Sistema EMEB es de ejecución automática, se debe presentar el resultado de la ejecución de los procesos. Así como también del registro de logs del propio sistema.

Se presentará también el resumen estadístico de la generación de bookmarks pero únicamente para el rol de Administrador del Wordpress multiusuario.

Capa Lógica de Negocios

En esta capa, se agrega los módulos principales del Sistema EMEB, como es la Obtención de metadatos y la Generación de Archivos RDF.

Cabe destacar que, se implementa el Web service Alchemy API para la extracción de etiquetas de las entradas que carecen de tags. Asi como también se utilizará el plugin Category Mapping para la categorización de los blogs basándose en una taxonomía de acuerdo a la división de areas de la universidad.

Capa de Datos

En esta división se encuentran, el nuevo esquema de base de datos e specificando que tablas se han utilizado para el funcionamiento del Sistema EMEB, e incluso el directorio donde son almacenados los RDF.

Con la finalidad de aprovechar la base de datos existente se propone hacer uso de dicha información mediante la agregación de una tabla wp_bookmarks, y la tabla wp_logs-bk la misma que constará de la extracción de la información necesaria para la generación del archivo RDF, a continuación se presenta su esquema modificado.

2.4.5 Estructura de la base de datos del WPMU

Wordpress multiusuario permite construir una comunidad sólida de blogs. Por ello la estructura de la plataforma de blogs, responde a la necesidad de administrar varios blogs de forma conjunta pero al mismo tiempo independiente uno del otro.

Cabe destacar que el wordpress multiusuario añade 8 tablas nuevas por cada blog nuevo, el nombre de las tablas se escriben con el prefijo wp_, seguido del número secuencial de creación del blog para distinguir las tablas de un blog determinado. Ejemplo: wp_2_comments, wp_2_links.

A continuación se presenta el esquema de las tablas creadas, básicas al momento de instalación y las tablas que son añadidas por cada nuevo blog. El # significa el número de identificador creado por cada blog.

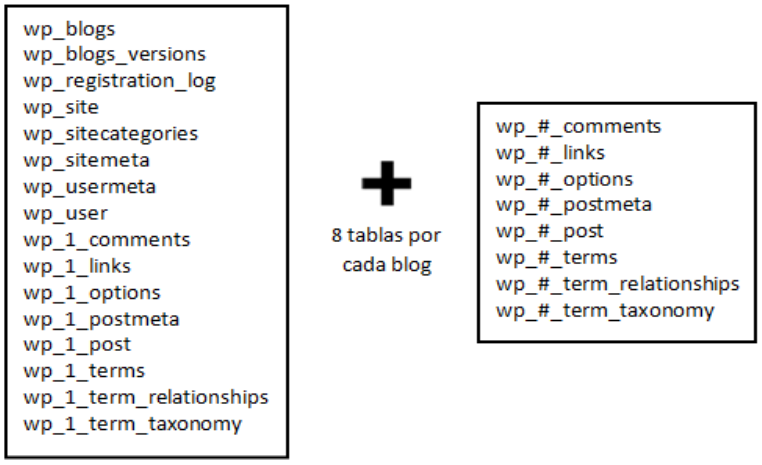


Figura 15. Tablas creadas por cada blog en el WPMU-UTPL

Según la Figura 16, se muestra las 17 tablas principales donde se almacena la información básica del WPMU.

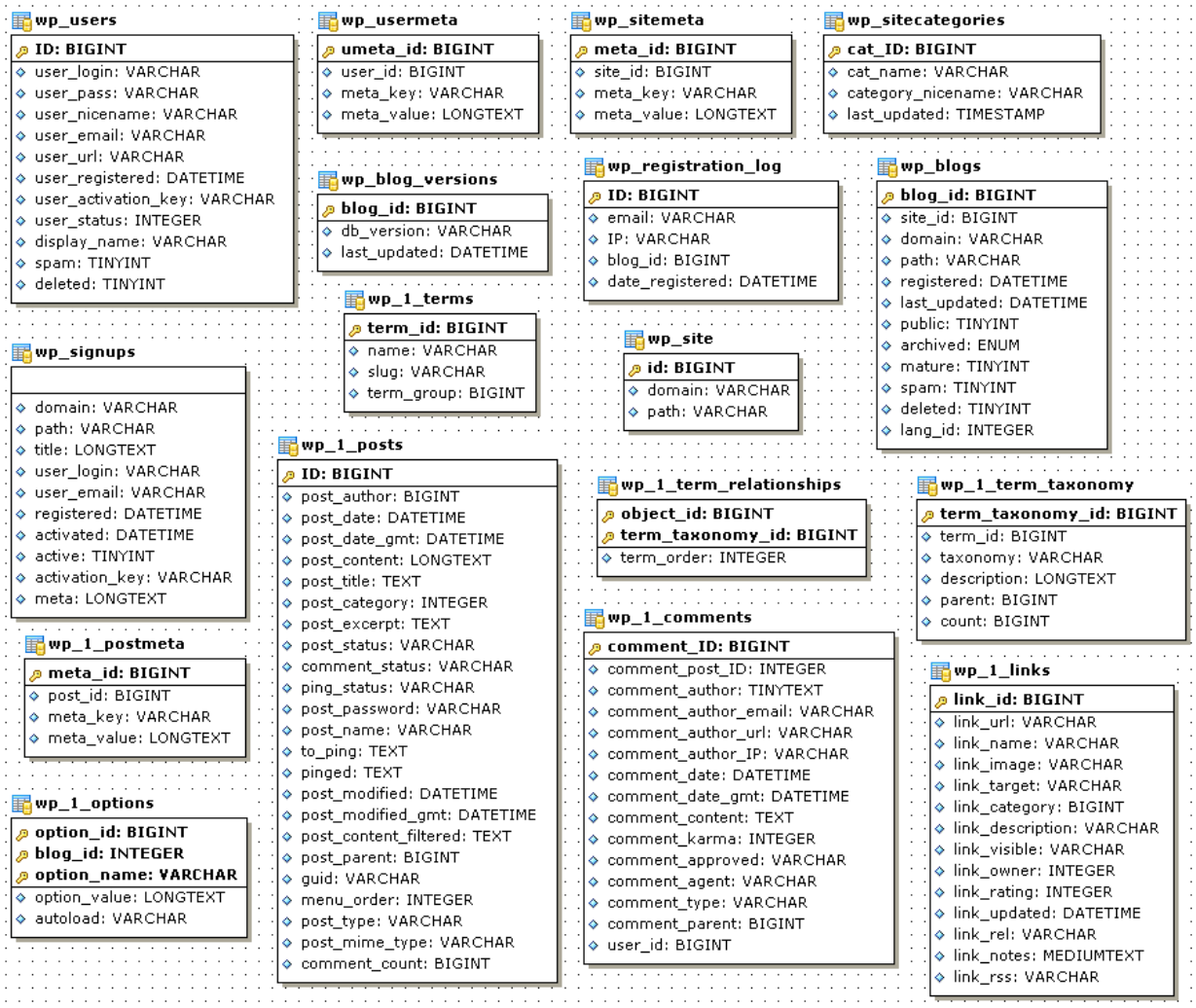


Figura 16. Esquema actual del WPMU versión 2.7.1 de la UTPL

Cada una de las tablas encontradas en la base de datos del WPMU posee información de los blogs cuya información es valiosa para la extracción de los datos necesarios para la descripción de recursos web.

2.4.6 Identificación de tablas del Wordpress necesarias para el desarrollo del Sistema

Dentro de la base de datos de WPMU se realiza la identificación de las tablas candidatas para la extracción de contenido.

Las características a tomar en cuenta para el desarrollo de un sistema de etiquetación a las entradas de cada uno de los blogs alojados en el WPMU-UTPL, se basan en la presentación como producto final que debe poseer las siguientes características:

- ✓ Nombre del Blog.
- ✓ Nombre de la entrada.

- ✓ Link de la entrada
- ✓ Autor.
- ✓ Descripción.
- ✓ Tags.
- ✓ Categorías

Tabla 9. Descripción de las tablas para la extracción de datos [11]

Nombre de la Tabla	Descripción
wp_blogs	Lugar donde se almacenan los id de todos los blogs creados en el sitio.
wp_user	La lista de usuarios se almacena es esta tabla.
wp_#posts	La información principal de WordPress son las entradas (**posts**), que son guardados en esta tabla.
wp_#terms	Las **categorías** de las entradas y links así como también las etiquetas de los post se encuentran en esta tabla.
wp_#term_relationships	Asociaciones de las entradas con categorías y etiquetas de la tabla wp_terms, junto con las asociaciones de links con sus respectivas categorías.
wp_#term_taxonomy	Contiene descripciones de la taxonomía (**categoría**, o **etiqueta**) para los datos mantenidos en la tabla wp_terms.
wp_#comments	Los comentarios en WordPress se mantienen aquí.
wp_#links	Información relacionada con las links.
wp_#options	Las opciones configuradas desde la administración en el WPMU, se puede extraer el nombre del blog de esta tabla.

La utilización de metadatos, es una solución para describir distintos objetos de información distribuidos en la web, de tal forma que, la búsqueda basada en estos metadatos sean utilizados en la extracción de información.

2.4.5 Esquema de Base de Datos propuesta para el Sistema EMEB

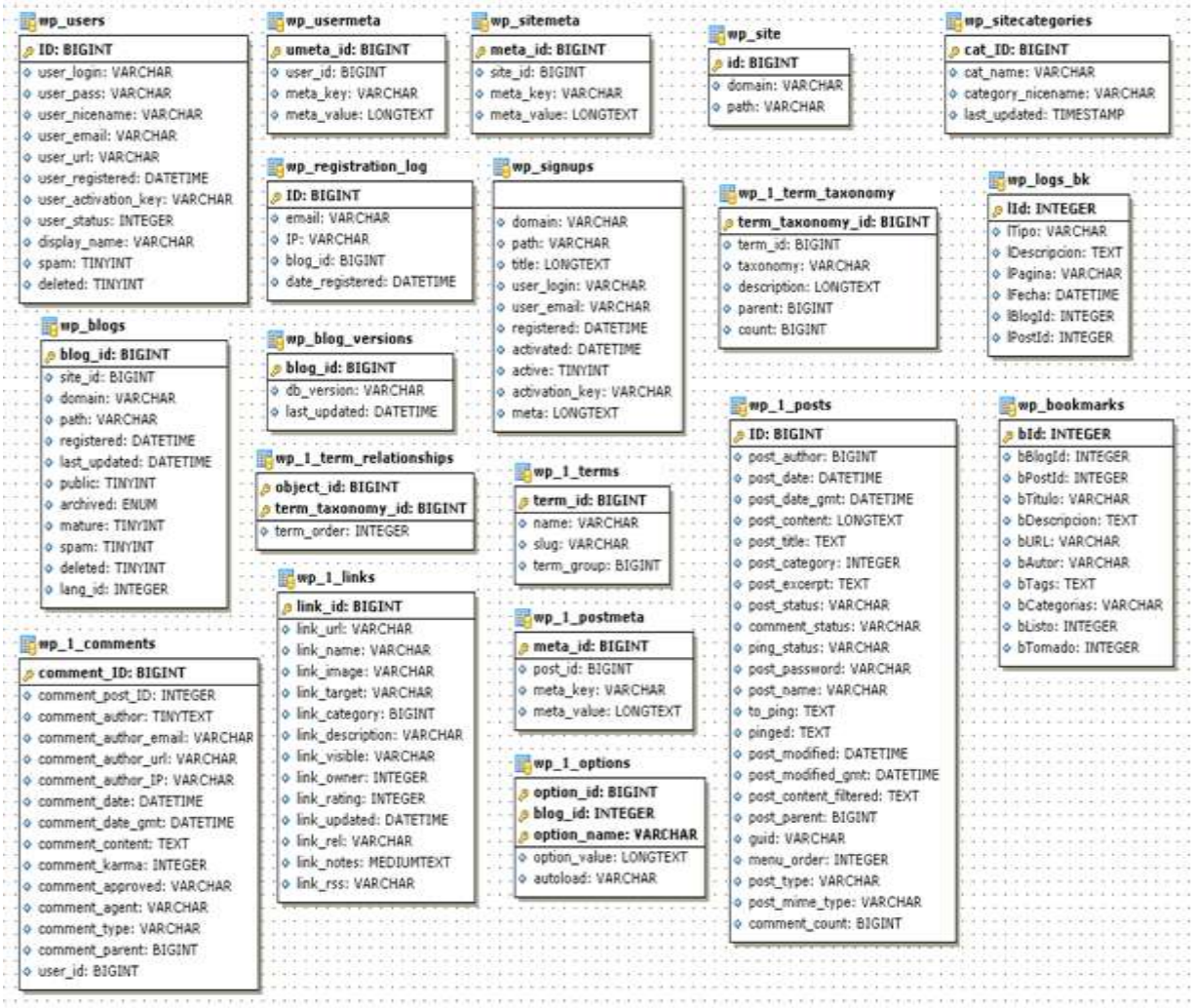


Figura 17. Esquema de Base de Datos Propuesta para el Sistema EMEB

Para la recolección de metadatos y generación del RDF de las entradas del WPMU, resulta más fácil la escritura de un archivo semántico desde una tabla que contenga toda la información para obtener archivos RDF sin errores y validados.

Las nuevas tablas que se agregan al wordpress multisuario UTPL, se explica en el *Anexo 13, literal 1*.

2.5 IMPLEMENTACIÓN DEL SISTEMA EMEB

El objetivo final de la implementación es generar archivos RDF que contengan la información de las entradas publicadas en los blogs del Wordpress, para lo cual tenemos dos procesos:

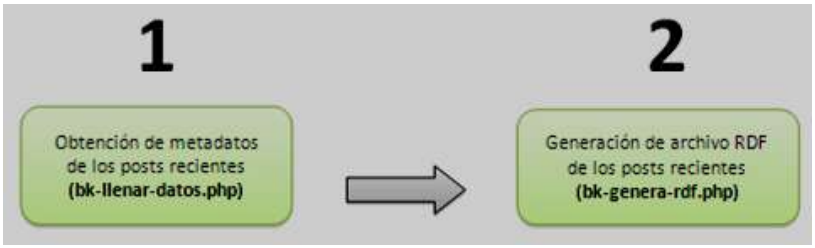


Figura 18. Pasos de la implementación del Sistema EMEB

2.5.1 Obtención de Metadatos

Para almacenar toda la información necesaria del sistema EMEB que se añaden dos tablas a la base de datos del WPMU, las mismas que están identificadas por el prefijo "wp_" seguido de los siguientes nombres:

- ✓ wp_bookmarks
- ✓ wp_logs

Estas tablas recogen la información necesaria para el funcionamiento del Sistema EMEB, como la información de cada bookmark y los logs del sistema que llevan un registro de la cantidad de bookmarks utilizados y toda la información necesaria para el administrador del wordpress multiusuario.

2.5.1.1 Proceso de la Obtención de Metadatos

Este proceso va ha buscar los post publicados por cada uno de los blogs de Wordpres MU, obtiene los metadatos disponibles y genera metadatos en caso de no existir y con esto crea un bookmark.

Detalle del proceso:

1. Obtener el listado de blogs disponibles.
2. Por cada Blog obtener el listado de post publicados.
3. Obtener la información de disponible de cada post.
4. En caso de faltar tags, mediante un proceso automático genera los tags para dicho post.
5. Con los datos obtenidos y generados inserta el bookmark en la tabla wp_bookmarks, de esta manera los bookmarks pueden ser utilizados por cualquier otro proceso.

2.5.2 Diagrama de Flujo de la Obtención de Metadatos

En el siguiente diagrama de flujo, se pueden visualizar los pasos que se deben seguir para la utilización del módulo de *Llenado de Datos*.

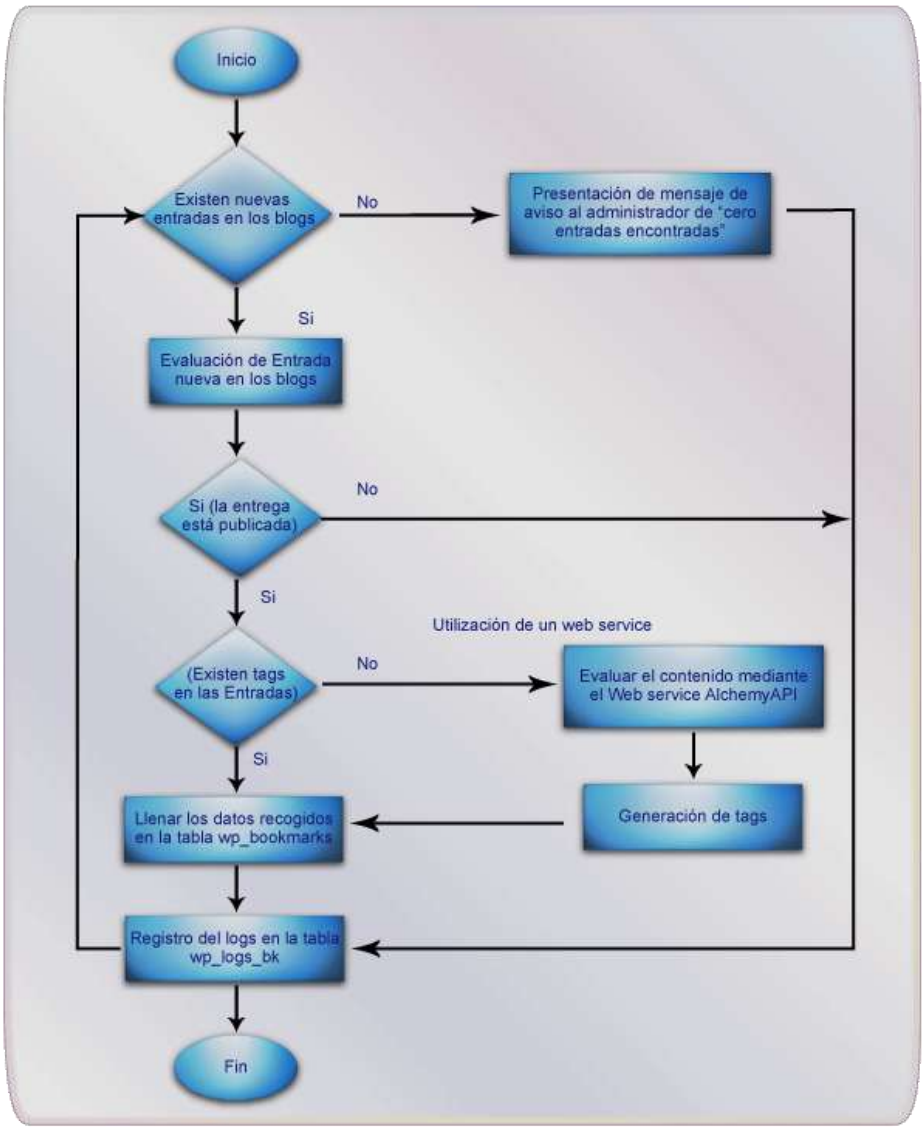


Figura 19. Diagrama de la Obtención de metadatos

1. Según el diagrama de flujo, se empieza con evaluación de las entradas que hayan sido ingresadas últimamente y no estén en la tabla wp_bookmark (según el campo bTomado (0,1))
2. En caso que no existiera entradas nuevas el sistema almacenaría un registro un que contendría: "Cero entradas encontradas y no se han llenado los datos".
3. En caso que si existieran entradas, se procede a la localización de la nueva entrada.
4. Seguidamente se evalúa si la entrada nueva tiene el estado "publicada".
5. Si la entrada tiene un estado diferente de "publicado" se registra en los logs y termina el proceso.
6. En caso que sea verdadero, es decir si la entrada tiene el estado "publicada" Se evalúa si la entrada posee etiquetas.
7. Si la entrada no tiene etiquetas, se utiliza el web serviceAlchemyAPI, pata llenar dicho información.

8. Luego se generan los tags para utilizarlos y almacenarlos en la tabla wp_bookmarks.(proceso alternativo)
9. En caso de que la entrada si posea etiquetas, se pasa al siguiente paso que es llenar los datos a la tabla wp_bookmarks
10. Todas las acciones se almacenan en los logs del sistema EMEB en la tabla wp_logs_bk.

Esquema de las tablas utilizadas para el llenado de datos

Por cada nuevo blogs se crean 8 tablas, para diferenciarse de las tablas principales del multiusuario, las mismas que se diferencian por un identificador único. Ejemplo:

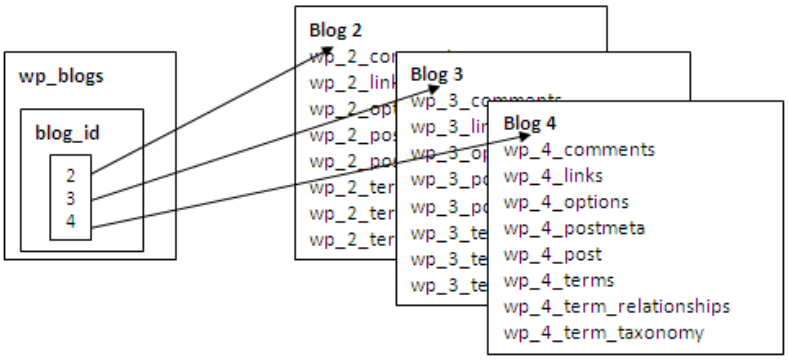


Figura 20. Distribución de blogs internos dentro del Wordpress Multiusuario

Para la extracción del nombre del blog se realiza el filtrado de datos desde la tabla wp-blogs, wp_#_post, wp_#_terms, wp_#_term_taxonomy, wp_1_cat_mapping, wp_#_term_relationships; donde # significa el numero de blog generado por cada creación de un blog. En la Figura 20, se muestra el esquema de las tablas del wordpress multiusuario de la UTPL, que se usan para el llenado de datos de la tabla wp_bookmarks.

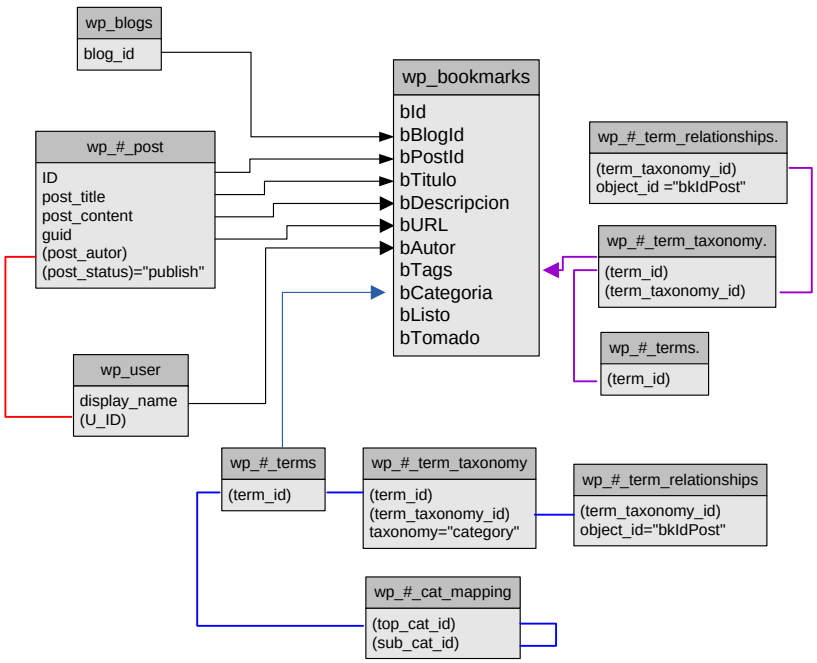


Figura 21. Esquema de las tablas utilizadas para la llenada de datos de la tabla wp_bookmarks

Donde: # significa el numero del blog de cada blog dentro del WMPU.

Las consultas que se realizan para la obtención de metadatos de las entradas de los blogs se describen en el *Anexo 13, literal 2*.

Se observa los datos de la tabla wp_bookmarks en la Figura 22.

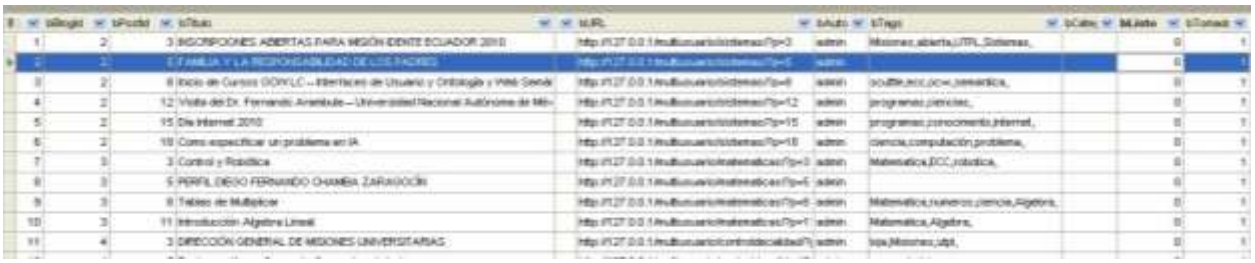


Figura 22. Carga de datos en la tabla wp_bookmarks

Restricciones para el llenado de datos:

- ✓ Un inconveniente es el formato de texto utilizado para la evaluación de contenido y posterior extracción de tags, evidenciando el inconveniente en el uso de tildes y ñ. *Ver Anexo 14, literal 1*.
- ✓ Para la obtención de un post se selecciona los 200 primeros caracteres del título del POST.
- ✓ Para la descripción obtiene los 200 primeros caracteres del POST.
- ✓ Debido a que en la descripción de algunas entradas tiene código HTML embebido en el cuerpo de la descripción, se realiza una limpieza de dichas etiquetas. *Ver Anexo 14, literal 2*
- ✓ Para la obtención de entradas nuevas se evalúa el máximo identificador del campo bPostId de la tabla wp_bookmarks. *Ver Anexo 14, literal 3*
- ✓ También se evalúan las entradas que no hayan sido tomadas anteriormente mediante la referencia de campo bTomado con los valores de 0 para entradas no generadas a RDF y 1 para aquellas ya incluidas en el Sistema EMEB. *Ver Anexo 14, literal 4*.
- ✓ Para la obtención de tags a partir de la generación del API, se toman hasta 120 caracteres.
- ✓ Para las categorías se realiza mediante el filtrado de la tabla wp_#_cat_mapping (añadida por el plugin Category Mapping).

2.5.3 Generar archivos RDF

Una vez generados los bookmarks se procede a la creación de un archivo RDF que los contenga, este archivo utiliza el estándar Dublín Core para la descripción de metadatos.

El proceso es el siguiente:

- ✓ Crea un archivo RDF cuyo nombre tendrá el siguiente formato: *bookmarks_[aaaammdd_hhmiss].rdf*
- ✓ Lee desde la tabla wp_bookmarks los bookmarks que aún no han sido tomados para la generación de archivos RDF.
- ✓ Por cada bookmark se genera la estructura en formato Dublín Core.

- ✓ Escribe el bookmark en formato DublinCore en el Archivo RDF.
- ✓ Una vez insertados todos los bookmarks se cierra el archivo rdf para se utilizado.
- ✓ El archivo RDF se deposita en un path (directorio) configurado por el usuario.

2.5.4 Diagrama de Flujo de Generar RDF

En el siguiente diagrama de flujo, se inicia después de la culminación del proceso de la *"Obtención de Datos";* en la siguiente figura se visualiza los pasos que se debe seguir para la utilización del módulo de *Generar RDF*.

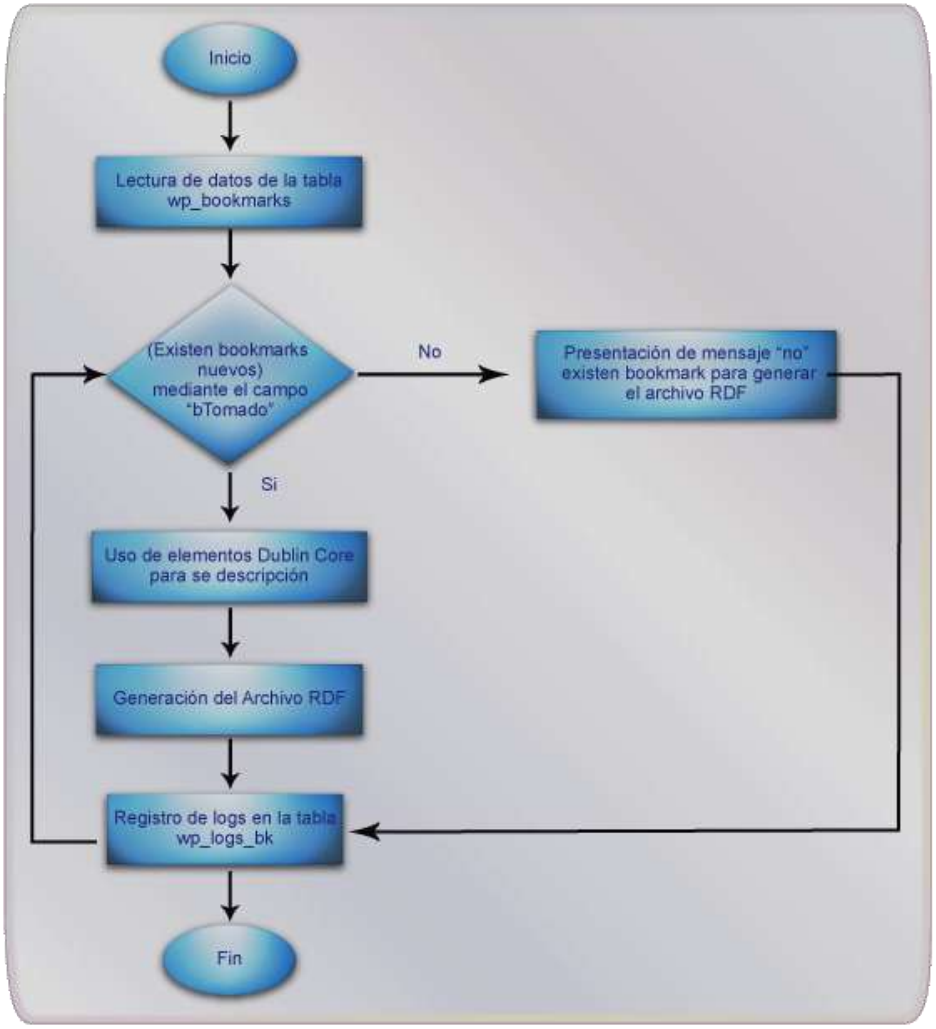


Figura 23. Diagrama de Flujo de Generar RDF

1. Se realiza la lectura de la tabla wp_bookmarks.
2. Se evalúa mediante el campo bTomado si ya ha sido tomado como bookmark, identificado por "0" cuando no han sido tomados, y "1" para aquellos que ya han sido generados.
3. En caso de no existir ninguna entrada nueva, se procede a registrar el mensaje de "no existen bookmarks para generar el RDF" finalizando el proceso.

4. Cuando si existan bookmarks nuevos se procede a utilizar dicha información utilizando la descripción de algunos de los elemento como: *dc:title, dc:description, dc:source, dc:creator, dc:subject de DublinCore.*
5. Se genera el archivo RDF y se almacena en el path *wordpressmu/bookmarks/bk-archivos-rdf.*
6. Finalizando con el registro de la actividad en la tabla wp_logs_bk, y en el resumen de logs que se genera por cada mes almacenado en la ruta *wordpressmu/bookmarks/log.*

2.6 CONFIGURACIÓN DEL SISTEMA EMEB

El Sistema EMEB contiene archivos y carpetas internas, de los cuales constan el procedimiento para su puesta en marcha y además de la configuración de archivos.

- ✓ Ruta de almacenamiento de archivos rdf. Ver anexo 15, literal 1
- ✓ Ejecución automática de los procesos. Ver anexo 15, literal 2
- ✓ Consulta de logs. Ver anexo 15, literal 3

3. En caso de no existir ninguna entrada nueva, se procede a registrar el mensaje de "no existen bookmarks para generar el RDF" finalizando el proceso.
4. Cuando si existan bookmarks nuevos se procede a utilizar dicha información utilizando la descripción de algunos de los elemento como: *dc:title, dc:description, dc:source, dc:creator, dc:subject de Dublin Core.*
5. Se genera el archivo RDF y se almacena en el path *wordpressmu/bookmarks/bk-archivos-rdf.*
6. Finalizando con el registro de la actividad en la tabla wp_logs_bk, y en el resumen de logs que se genera por cada mes almacenado en la ruta *wordpressmu/bookmarks/log.*

2.6.1. Instalación y administración del plugín Bookmarks en RDFs

Se desarrolla el plugín Bookmarks en RDFs, el mismo que le permite al administrador mantenerse informado de los blogs más activos, así como también el resumen de los post que han sido generados a RDF. El proceso de instalación y uso se detalla en el *Anexo 16*.

CAPITULO 3

3. PLAN DE VALIDACIÓN Y PRUEBAS

En el presente capitulo se detalla la fase de validación y pruebas de la adaptación del sistema EMEB, para determinar los resultados obtenidos y proceder a concluir, se procede a realizar las pruebas de: integridad de datos, funcionamiento, validación, accesibilidad y estabilidad.

3.1 INTRODUCCIÓN

3.1.1 Propósito

El presente plan tiene como propósito llevar a cabo una estrategia de pruebas y validaciones, con la finalidad de reducir al mínimo los riesgos debido a fallas operativas así como también asegurar el objetivo planteado en el inicio del proyecto.

3.1.2 Objetivos

La elaboración del plan de pruebas tiene como objetivos:

- ✓ Evaluar la confiabilidad, funcionalidad del sistema EMEB.
- ✓ Identificación de errores encontrados posterior a la adaptación del sistema EMEB al Wordpress multiusuario de la UTPL.
- ✓ Puntualizar las estrategias de prueba a ser empleadas

3.1.3 Audiencia

La audiencia utilizada para el presente proyecto es:

- ✓ Administrador o desarrollador del proyecto.
- ✓ Usuarios finales (profesores, estudiantes, personal que interactúan con los blogs alojados en la plataforma multiusuario).

3.2 PLAN DE PRUEBAS

Para la aplicación del presente plan de pruebas, se ha configurado un subdominio en el cual se realiza la simulación con la base de datos, los cuales tienen los permisos de: lectura y escritura para el administrador del wordpress.

Una vez realizadas las configuraciones y la adaptación del sistema EMEB-UTPL, al wordpress multiusuario de la universidad, para ello se proponen las siguientes pruebas a nivel de sistema y nivel de usuario.

Los perfiles de usuarios considerados para los tipos de pruebas se presentan en la Tabla 10.

Tabla 10. Tipos de prueba por perfil de usuario

Perfil de usuario	Tipo de Prueba
Usuario administrador	Pruebas de integridad de datos Pruebas de funcionamiento Pruebas de validación

	Pruebas de accesibilidad
Usuario registrado (interactúa con los blogs)	Pruebas de estabilidad
Usuario auditor	Pruebas de precisión

3.2.1. Pruebas de Integridad de Datos

Objetivo:

Asegurar que la extracción de datos desde las tablas utilizadas de la base de datos funcione correctamente.

Técnica:

Revisar la base de datos para asegurar que los datos han sido cargados automáticamente en las tablas añadidas por el sistema EMEB y que estén llenos todos los campos devueltos a guardarse en la tabla donde se almacena dicha información.

Población:

Las pruebas se realizan por el administrador y constan de dos partes:

- ✓ Realización de las operaciones disponibles en el Sistema EMEB
- ✓ Utilización de los formatos (escenarios de prueba) para la presentación de los resultados.

Escenarios de prueba

En esta prueba es necesario especificar que existen dos situaciones que se presentan en las entradas de los blogs del multiusuario:

1. Llenado de Datos en la tablas del Sistema EMEB
    a. Cuando existen tags asociadas a las entradas.
    b. Cuando no existen tags en las entradas
2. Generación de Archivos RDF

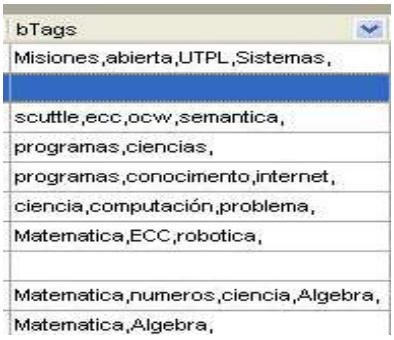
Escenario 1.a: Llenado de Datos en las tablas del Sistema EMEB cuando existen tags asociadas a las entradas.

Resultado:

Una vez ejecutado el método de llenado de datos se evidencia que son datos correctos alojados en la tabla *wp_bookmarks*, sin interrupción de alguno de ellos.

Escenario 1.b: Llenado de Datos en las tablas del Sistema EMEB cuando no existen tags asociadas a las entradas

En éste caso (cuando no existan tags en las entradas), se procede a la verificación que se haya utilizado el web service y se almacene en la tabla wp_bookmarks en la columna bTags, como se muestra en la Figura 24.



bTags
Misiones,abierta,UTPL,Sistemas,
scuttle,ecc,ocw,semantica,
programas,ciencias,
programas,conocimento,internet,
ciencia,computación,problema,
Matematica,ECC,robotica,
Matematica,numeros,ciencia,Algebra,
Matematica,Algebra,

Figura 24. Columna bTags de la tabla wp_bookmarks

Resultado:

Los tags extraídos mediante el Web Service Alchemy API, fueron utilizados como parte de la descripción de cada bookmarks, referenciándose a una entrada de determinado blog.

Escenario 2: Generación de Archivos RDF

En éste escenario se prueba la funcionalidad de la generación de archivos rdf que describen los bookmarks de cada una de las entradas nuevas que se alojen en los blogs del wordpress multiusuario de la UTPL.

Además del almacenamiento de archivos RDF generados en la ruta especificada como se muestra en la Figura 25, almacenados con el prefijo "bookmarks_" seguido por la fecha de creación, dichos archivos serán utilizados como puente entre los sistemas EMEB y aplicaciones de representación de bookmarks

Nombre	Tamaño	Tipo	Fecha de modificación
bookmarks_20100712_065306	6 KB	Archivo RDF	12/07/2010 1:57
bookmarks_20100712_103632	6 KB	Archivo RDF	12/07/2010 5:36
bookmarks_20100712_104044	6 KB	Archivo RDF	12/07/2010 5:40

Figura 25. Almacenamiento de archivos RDF generados por el sistema EMEB

3.2.2 Pruebas de Funcionamiento:

Objetivo:

Asegurar el cumplimiento de la funcionalidad de la navegación, carga de datos, procesamiento, y generación de archivos RDFs.

Técnica:

Se ejecutan los procesos automáticos, usando una copia de la base de datos del wordpress actual, al sitio de pruebas cuyo link es: http://blogsprueba.utpl.edu.ec , en cuyos resultados se verifica lo siguiente:

- ✓ Qué los procesos automáticos se ejecuten según el cronograma diario que se configuró en el servidor. *Ver anexo 14, literal 2.*
- ✓ Qué los resultados esperados ocurran cuando se usen datos válidos.

- ✓ Que se registren en los logs del sistema mensajes apropiados de error y precaución cuando no existan entradas nuevas.

Resultado:

- ✓ Todas las pruebas planeadas han sido ejecutadas.
- ✓ Todas las irregularidades identificadas han sido corregidos.*Ver anexo 13, literales 1,2,3 y 4.*

En el proceso de las pruebas de funcionamiento, se verificó que la extracción de datos, carga y generación de los procesos del sistema EMEB se ejecutan correctamente.

3.2.3 Pruebas de Validación:

Este tipo de pruebas es necesario realizar una validación del archivo RDF.

Objetivo:

Certificar que los archivos RDF generados por el sistema EMEB cumplan con las normas establecidas para la utilización de la información que describe dicho RDF.

Técnica:

Para la evaluación de los RDF se utiliza el validador online de la W3C, en el cuál se puede obtener el gráfico del modelo y sus tripletas.

```
  </rdf:Description>
- <rdf:Description rdf:about="http://blogsprueba.utpl.edu.ec/websematica/?p=6">
    <dc:title>Drupal y la Web Semántica</dc:title>
    <dc:description>Hace mucho tiempo se viene hablando de la tan anunciada Web semantica sin
      embargo hasta hoy solo es un proyecto que no llega a solidificarse, a pesar de esto muchos
      sitios estan optando por tecnologia...</dc:description>
    <dc:source>http://blogsprueba.utpl.edu.ec/websematica/?p=6</dc:source>
    <dc:creator>mlortiz5</dc:creator>
    <dc:identifier>Area Técnica</dc:identifier>
    <dc:subject>web semántica</dc:subject>
    <dc:subject>portal universitario</dc:subject>
    <dc:subject>datos vía rdfa</dc:subject>
    <dc:subject>Drupal 6.x</dc:subject>
    <dc:subject>CMS Drupal</dc:subject>
    <dc:subject>Web semantica</dc:subject>
    <dc:subject>Dries Buytaert</dc:subject>
    <dc:subject>módulo rdf</dc:subject>
    <dc:subject>formato rdf</dc:subject>
    <dc:subject>módulo evoc</dc:subject>
  </rdf:Description>
```

Figura 26. RDF que describe a un bookmark

Con el desarrollo de estas pruebas de validación se obtuvo un conjunto de errores en base a la información en el campo de *description,* por el motivo que existe código embebido de sitios de redes sociales como slideshare, youtube, flickr y similares. Esto daba como resultado un error al momento de la validación y posterior visualización del grafo.

Tabla 11. Casos de prueba de validación

Casos de Prueba	N°	%
No generan error	30	93,33
Genera error	2	6,67
TOTAL	32	100,00

Se encontró que el 6,67% de los casos de prueba, para distintas funcionalidad, se produjeron errores en la descripción.

En la siguiente figura se muestra la relación entre el número de casos de prueba frente al número de casos de prueba que generaron error.

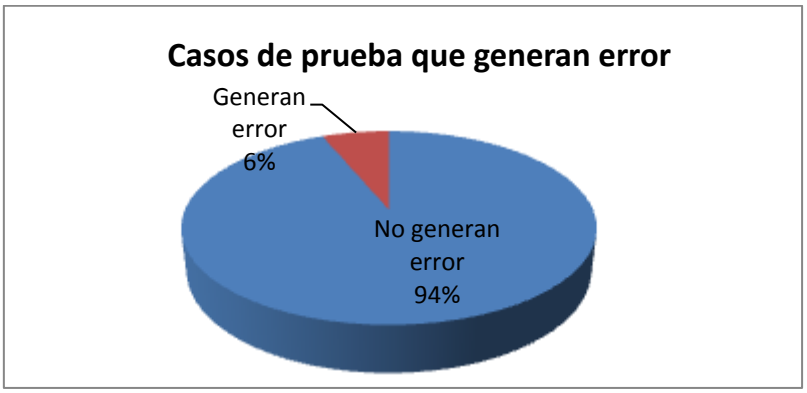


Figura 27. Casos de prueba que generan error

Para la mitigación de este inconveniente se procedió a utilizar una función de limpieza de código HTML, con la finalidad de depurar datos no entendibles y el tratamiento para que la información de la descripción tenga validez visual.

Una vez corregido este error se procedió a la validación del código para la obtención de tripletas y el grafo del RDF evaluado.

Cabe destacar que, cuando una entrada no posea todos los metadatos completos, el Sistema EMEB genera la descripción del bookmark con los datos que ha sido posible extraer desde la base de datos del Wordpress multiusario. De esta manera no se limita la funcionalidad, o ámbito de satisfacción del desarrollo. Esta carencia de metadatos en especial en la descripción se debe a:

1. En caso que el usuario ingresó una entrada con un solo archivo, y no colocó una descripción en el área de contenido, el campo de descripción se quedará vació, siendo estructurado en el RDF con los demás metadatos.
2. En caso que el usuario ingresó un archivo y no colocó tags en la entrada, el Sistema EMEB no podrá hacer uso de AlchemyAPI para la extracción de etiquetas por la restricción que no poseer contenido que analizar, por lo tanto estos campos almacenarán con datos vacios.

En cuanto a la validación del RDF, su resultado fue satisfactorio; lo que significa que pasa esta prueba con cero errores.

Triples of the Data Model

Number	Subject	Predicate	Object
1	http://blogaprueba.utpl.edu.ec/webasematica/?p=6	http://purl.org/dc/elements/1.1/title	"Drupal y la Web Semántica"
2	http://blogaprueba.utpl.edu.ec/webasematica/?p=6	http://purl.org/dc/elements/1.1/description	"Hace mucho tiempo se viene hablando de la tan anunciada Web semantica sin embargo hasta hoy solo es un proyecto que no llega a solidificarse, a pesar de esto muchos sitios estan optando por tecnologia..."
3	http://blogaprueba.utpl.edu.ec/webasematica/?p=6	http://purl.org/dc/elements/1.1/source	"http://blogaprueba.utpl.edu.ec/webasematica/?p=6"
4	http://blogaprueba.utpl.edu.ec/webasematica/?p=6	http://purl.org/dc/elements/1.1/creator	"mlortiz5"
5	http://blogaprueba.utpl.edu.ec/webasematica/?p=6	http://purl.org/dc/elements/1.1/identifier	"Area TÃ©cnica"
6	http://blogaprueba.utpl.edu.ec/webasematica/?p=6	http://purl.org/dc/elements/1.1/subject	"web semántica"
7	http://blogaprueba.utpl.edu.ec/webasematica/?p=6	http://purl.org/dc/elements/1.1/subject	"portal universitario"
8	http://blogaprueba.utpl.edu.ec/webasematica/?p=6	http://purl.org/dc/elements/1.1/subject	"datos via rdfa"
9	http://blogaprueba.utpl.edu.ec/webasematica/?p=6	http://purl.org/dc/elements/1.1/subject	"Drupal 6.x"
10	http://blogaprueba.utpl.edu.ec/webasematica/?p=6	http://purl.org/dc/elements/1.1/subject	"CMS Drupal"
11	http://blogaprueba.utpl.edu.ec/webasematica/?p=6	http://purl.org/dc/elements/1.1/subject	"Web semantica"
12	http://blogaprueba.utpl.edu.ec/webasematica/?p=6	http://purl.org/dc/elements/1.1/subject	"Dries Buytaert"
13	http://blogaprueba.utpl.edu.ec/webasematica/?p=6	http://purl.org/dc/elements/1.1/subject	"mÃ³dulo rdf"
14	http://blogaprueba.utpl.edu.ec/webasematica/?p=6	http://purl.org/dc/elements/1.1/subject	"formato rdf"
15	http://blogaprueba.utpl.edu.ec/webasematica/?p=6	http://purl.org/dc/elements/1.1/subject	"mÃ³dulo evoc"

Figura 28. Representación del RDF evaluado en Tripletas

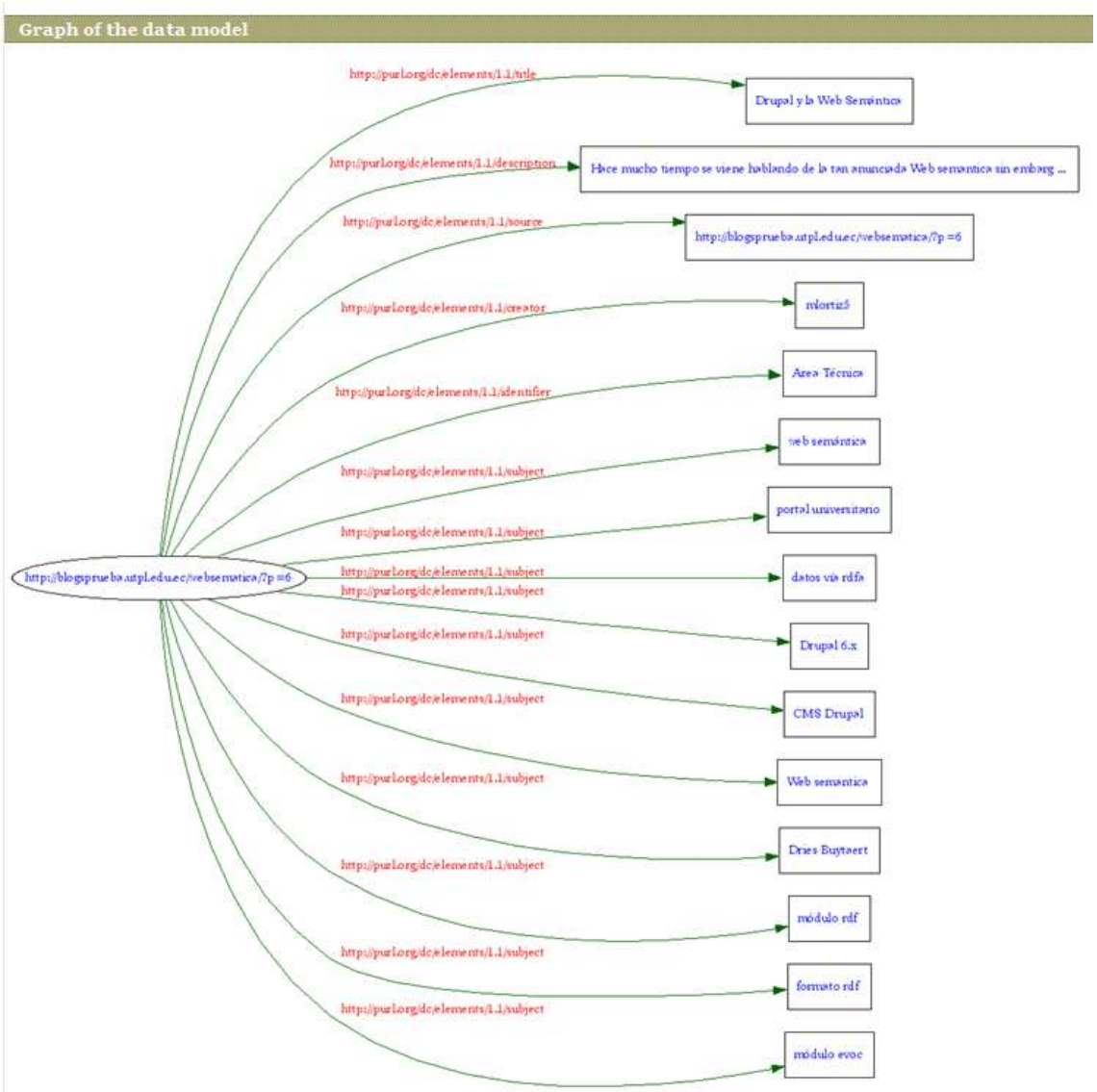


Figura 29. Representación en árbol del RDF evaluado

Resultado:

Se deduce que el RDF se no contine ningún tipo de error repostado en la validación, es decir; estos archivos pueden ser empleados en cualquier sistema de representación de bookmarks.

3.2.4 Pruebas de Accesibilidad

La estructura de los RDFs generados se pueden visualizar a través de algunos proyectos, navegadores y exploradores desarrollados por organizaciones y/o Universidades tales como Ping The Semantic Web, Sindice, ZitGist y otros, comprobando inferencias de sus clases y propiedades.

Mediante el sitio Ping The Semantic Web, se registró algunos RDF generados por el Sistema EMEB.

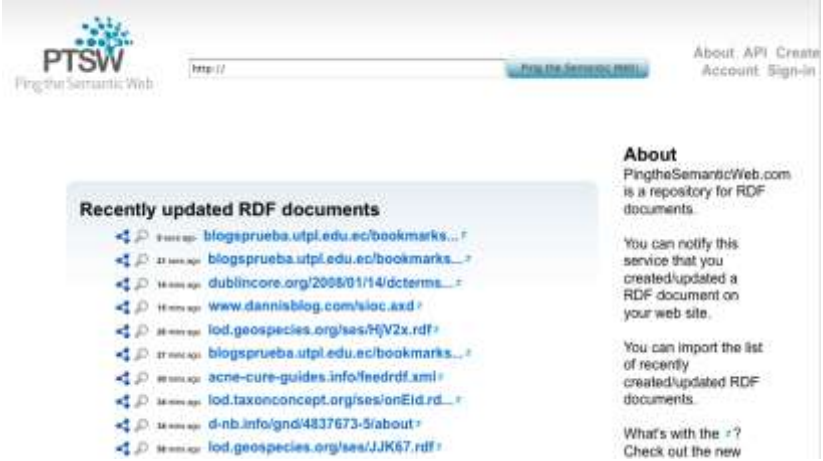


Figura 30. Resultados obtenidos en Ping The Semantic Web (http://pingthesemanticweb.com/)

Ping The Semantic Web es como un repositorio de enlaces hacia documentos RDF y se visualiza mediante la herramienta Zitgist donde se obtiene los siguientes resultados:

Figura 31. Resultados obtenidos en ZitGist (www.dataview.zitgist.com)

Zitgist ofrece un visualizador de datos RDF (dataviewer) el mismo que muestra la estructura con la cual está organizado el documento RDF y presenta información relativa al mismo.

Mediante el Proyecto Sindice desarrollado por DERI (Digital Enterprise Research Institute www.deri.ie) se obtiene los siguientes resultados:

Figura 32. Resultados obtenidos en The Semantic Web Index (http://sindice.com/search?q=historia+utpl&qt=term)

Sindice permite suscribir las direcciones de los RDFs además de realizar búsqueda de los mismos con el texto deseado presentando también su información desglosada.

3.2.5 Pruebas de Estabilidad:

Objetivo:

Asegurar que el sistema EMEB es estable y no altera el funcionamiento del Wordpress multiusuario, y además los casos de prueba permitirá probar todas las funcionalidades del sistema.

Técnica:

Almacenar entradas de blogs en todos los casos de pruebas para asegurar su estabilidad.

Población:

Las pruebas se realizan mediante un test realizado a los usuarios con la finalidad de ingresar entradas (post), con los siguientes casos de prueba: *Ver anexo 17*.

Tabla 12. Casos de prueba de estabilidad

Casos de prueba	Número de Entradas
Entrada de texto con etiquetas	5
Entrada de texto sin etiquetas	5
Entrada con archivo y etiquetas	4
Entrada con archivo y sin etiquetas	4
Entradas sin contenido	5
Entradas sin categorías	5
Entradas con codigo embebido	4
Total de entradas	32

Con respecto a la segunda opción del test (*Anexo 17)* cuyo enunciado es: "La entrada que ud. ingreso se publico normalmente".Según el resultado de esta pregunta, se obtiene la siguiente gráfica.

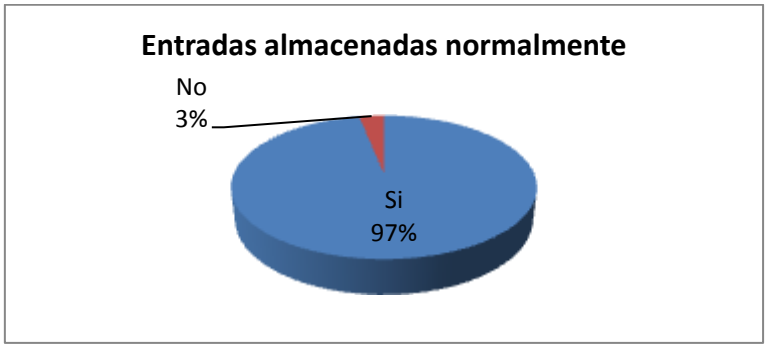


Figura 33. Pruebas de estabilidad

Las entradas que se almacenaron normalmente reflejan el 97%, con respecto al 3% de un caso en donde el archivo era demasiado grande y sopresaba el límite de peso normal de subida de archivos hacia la plataforma del WPMU-UTPL.

Resultado:

Durante esta evaluación se observó que se almacenan las entradas de manera normal. Por lo tanto el Sistema EMEB se considera estable y no invasivo en el funcionamiento normal del WPMU-UTPL.

3.2.6 Pruebas de Precisión:

Objetivo:

Evaluar que los tags ingresados por el usuario estén considerados en los tags extraídos por Alchemy Api.

Técnica:

Para realizar esta prueba se utiliza la fórmula para la extracción de muestras que se detalla a continuación:

$$n = \frac{Z^2 * P * Q * N}{E^2(N-1) + Z^2 * P * Q}$$

Cuya nomenclatura es:
n = Número de elementos de la muestra
N = Número de elementos de la población o universo (existen 1174 entradas)
P/Q = Probabilidades con las que se presenta el fenómeno
$\mathbf{Z^2}$ = Valor crítico correspondiente al nivel de confianza elegido; siempre se opera con valor zeta 2, luego Z = 2.
E = Margen de error permitido (determinado por el responsable del estudio)
Porcentaje de certeza del 90%

Margen de error del 10%

$$n = \frac{1{,}96^2 * 50 * 50 * 1174}{10^2(1174 - 1) + 1{,}96^2 * 50 * 50}$$

$$n = 88{,}8474437$$

De la muestra obtenida se analiza 89 entradas en el blog multiusuario de la UTPL. Se toma en cuenta los siguientes aspectos:

- ✓ Que posean tags ingresados por el usuario.
- ✓ Que contengan un mínimo de texto para ser evaluados por Alchemy Api.

Luego de la extracción manual de las 89 entradas etiquetadas se descubrieron 260 tags, para posteriormente evaluarlas con Alchemy Api y extraer tags cuyo resultado es de 625 tags.

Para realizar este test, se utiliza la puntuación de F1.

1. Definición de F1 score

En estadísticas, la puntuación F_1 es una medida de la prueba de precisión, se tiene en cuenta la precisión p y el recall r, que se utiliza para calcular la puntuación F1. La puntuación de la F_1 se puede interpretar como una media ponderada de la precisión y recall, donde un F alcanza su mejor valor a 1 y la peor puntuación a 0.

En la recuperación de metadatos, la precisión se puede medir de diferentes maneras, como las medidas utilizadas por la TREC[16].

La hipótesis para esta prueba es *"Los tags del usuario están inmersos en los tags extraídos por Alchemy"*

Tabla 13. Distribuciones de clasificación [12]

		resultado correcto / clasificación	
		Los tags del usuario están inmersos en los tags extraídos por Alchemy	***Los tags del usuario no están inmersos en los tags extraídos por Alchemy***
resultado obtenido/ clasificación	***Los tags del usuario están inmersos en los tags extraídos por Alchemy***	***VP:*** *verdaderos positivos*	***VN:*** *verdaderos negativos*
	Los tags del usuario no están inmersos en los tags extraídos por Alchemy	***FP:*** *falsos positivos*	***FN:*** *falsos negativos*

Precisión es la fracción de tags recuperados que son notables como respuesta a una evaluación de determinado texto.

$$Precisión = \frac{VP}{VP + FP}$$

[16] TREC (Enfoques de la alta exactitud de recuperación), se inició en 1992, co-patrocinado por el NIST y el Departamento de Defensa de EE.UU., con el propósito de apoyar la investigación dentro de la comunidad de recuperación de información, proporcionando la infraestructura necesaria para la evaluación a gran escala de las metodologías de recuperación de texto.

Recall es el porcentaje de tags recuperados, en realidad responde la extracción de determinado texto. En otras palabras, consiste en todos los tags de respuesta que debería haberse producido con una extracción.

$$Recall = \frac{VP}{VP + FN}$$

La puntuación de la *F1* es una combinación de la recuperación y precisión, proporcionando una medida de exactitud total.

$$F1\ score\ = 2 * \frac{Precision * Recall}{Precision + Recall}$$

El nivel de exactitud es la proporción de resultados verdaderos y se evalúa con la formula.

$$Exactitud = \frac{(VP + VN)}{(VP + VN + FP + FN)}$$

2. Tratamiento de los Tags

El idioma español es muy extenso y existen varias palabras que son utilizadas con mucha frecuencia, estas palabras no son consideradas por ningún buscador, sino que son filtradas quedando fuera de cualquier indexación [13].

Cuando se trabaja con datos es muy común encontrar similitudes de palabras (tags) debido a errores de tipeo, etc. ocasionando alteraciones de caracteres entre palabras. Para este caso se aplica el algoritmo de la Distancia de Levenshtein [14], que consiste en realizar el mínimo de operaciones requeridas para transformar una cadena de caracteres en otra. En la comparación de la similitud entre cadenas de caracteres. Ejemplo:

- ✓ memoria dinámica
- ✓ memoria ram (umbral de 8 caracteres)

En ambos experimentos de evalúan 260 tags ingresados por usuarios, frente a 625 tags extraídos por Alchemy de las mismas entradas. *Ver Anexo 18, literal 1 y 2.*

3. Experimento 1: *Aplicando el Algoritmo de la Distancia de Levenshtein.*

En el tratamiento de los tags se utiliza un umbral cuyo promedio es de 5 caracteres. *Ver anexo 18, literal 3.*

El cuadro de los datos obtenidos se manifiestan en el *Anexo 18, literal 4.*

Tabla 14. Resultados obtenidos

Precisión	Recall	F1_score	Exactitud
0,95555556	0,87755102	0,91489362	0,9704797
95,5%	87,7%	91,4%	97%

Después de revisar manualmente los tags se obtiene una precisión del 95,5%, un recall de 87,7% y el valor de F1_score es de 91,4 % respondiendo afirmativamente a la hipótesis "*Los tags del usuario están inmersos en los tags extraídos por Alchemy*" y en

efecto responde a la expectativa de exactitud de la hipotesis. Esto sugiere que es "suficientemente aceptable", con una exactitud de los datos analizados del 97%.

4. **Experimento 2: *Sin aplicación de algoritmos.***

Los datos obtenidos se explican en el *Anexo 18, literal 5*.

En esta prueba no se aplica ningún algoritmo y se obtiene los siguientes resultados en medidas de precisión:

Tabla 15. Resultados obtenidos

Precisión	Recall	F1_score	Exactitud
0,92	0,79310345	0,85185185	0,97003745
92%	79,3%	85,1%	97%

Los resultados obtenidos tienen el valor del 92% de precisión, cuyo recall es de 79,3%, proyectado la *puntuación F1* de 85,1% es "aceptable", con una exactitud de los datos analizados del 97%, por lo tanto se acepta la hipótesis "*Los tags del usuario están inmersos en los tags extraídos por Alchemy*".

Resultados:

Aplicando el algoritmo de la Distancia de Levenshtein	Sin aplicación de algoritmos
✓ La precisión de los tags extraídos refleja un valor de 95,5%, y significa que los tags ingresados por los usuarios están inmerosos en los tags que extrae Alchemy. ✓ El recall significa que los tags ingresados por los usuarios se asemeja en el 87,7/% al grupo de tags que se extraería con Alchemy API. ✓ El valor de F1 es de 91,4% lo que significa que es suficientemente bueno aplicando el algoritmo de la Distancia de Levenshtein. ✓ La exactitud es de 97% y significa que se ha reconocido todos los actuales positivos.	✓ La precisión de los tags extraídos tiene un valor de 92%, y significa que que los tags ingresados por los usuarios están inmerosos en los tags que extrae Alchemy. ✓ El recall tiene un valor de 79,3% de probabilidad que contenga los mismos tags ingresados por los usuarios frente a los tags que se extraería por Alchemy API. ✓ El valor de F1, sin aplicar ningún algoritmo de acoplación de caracteres, su valor es de 85,1% cuyo resultado "aceptable" en su funcionamiento. ✓ La exactitud es de 97% y significa que se ha reconocido todos los actuales positivos.

En sistemas de recuperación de información se busca que los tags sean en lo posible similares a los ingresados por humanos, ya que la tecnología aun no supera al pensamiento del cerebro humano.

Cuando se necesite de sistemas óptimos de etiquetación, Alchemy Api resuelve en parte dicho problema, pero se debe considerar el tratamiento a los tags extraídos, como; eliminar las palabras que no son tomadas en cuenta por la mayoría de buscadores, además utilizar algoritmos para conseguir similitudes de tags, también adaptar aplicaciones de lexemas propias del idioma o algoritmos de reducción de la dimensionalidad.

4: DISCUSIÓN

El proceso basado a la solución de etiquetación de contenidos de blogs dentro del blog multiusuario de la UTPL, el mismo que en este proceso se realiza y aplica una taxonomía con la finalidad de categorizar los blogs y contribuir a la organización de acuerdo la ideología de distribución de la UTPL, y por otro lado la información contenida en cada uno de los blogs suele ser importante y aporta al conocimiento de la sociedad, en este trabajo se extrae la información más significativa de cada entrada de todos los blogs en manera de bookmarks que utilizando la descripción semántica y el estándar Dublín Core en un archivo rdf de calidad, y que a su vez se almacena en un dominio público donde esta información pueda ser utilizada por repositorios de representación de contenido, logrando así reducir la brecha entre el conocimiento casi invisible entre la información de los blogs y la comunidad.

Al finalizar las pruebas se deduce que la aplicación de Alchemy API ayuda en parte a la carencia de etiquetación de entradas de blogs publicados por usuarios, y aún no es posible aproximarse a la inteligencia del cerebro humano para clasificar y proporcionar una etiquetación de excelencia, es por esta razón que se debería difundir la idea de agregar etiquetas a recursos educativos en cualquier sitio web, más aún si esta tiene carácter educativo.

5. CONCLUSIONES

Luego de haber culminado el desarrollo del proyecto de fin de carrera: " Desarrollo de un sistema semiautomático de etiquetación de contenidos de blogs utilizando lenguajes de descripción semántica", se ha llegado a establecer las siguientes conclusiones:

- ✓ Del análisis realizado en los contenidos publicados en la plataforma WPMU de la UTPL se encuentra que, el 62% de las entradas carece de etiquetación; además el 19% de los blogs no presenta categorización de sus contenidos de ningún tipo.
- ✓ El uso de web services como el AlchemyApi permite la extracción de palabras clave de las entradas del Wordpress Multiusuario, solucionando el problema de falta de etiquetado por parte de los usuarios.
- ✓ El Sistema EMEB permite la extracción de palabras claves, desde páginas HTML sin depender del contexto, con la limitante que debe existir por lo menos 200 caracteres.
- ✓ El uso de Dublín Core en la generación de RDF ha permitido la estandarización de bookmarks que genera el sistema EMEB, facilitando la difusión de los mismos hacia plataformas que manejan este estándar.
- ✓ La taxonomía definida para los blogs del wordpress multiusuario de la UTPL permitió la categorización de los mismos de acuerdo a la estructura organizacional de la universidad.
- ✓ El sistema EMEB mediante la obtención de archivos RDF, permitirá que los contenidos de los blogs de la UTPL, puedan ser registrados en buscadores semánticos y utilizados por herramientas de representación semántica.
- ✓ En particular, después de examinar la relevancia de la extracción de tags realizado por el Sistema EMEB, se obtuvo un 85,1% mediante la aplicación de F1 score en medidas de precisión, lo que significa que su funcionamiento es suficientemente aceptable.

6. RECOMENDACIONES

- ✓ Antes de implementar el Sistema EMEB se recomienda realizar una depuración de la base de datos del WPMU-UTPL, actualmente existen 13592 tablas, de las cuales la mayoría no contienen datos y pertenecen a los blogs que están inactivos.
- ✓ Incentivar la formación de una cultura de etiquetación de contenidos en plataformas web, con la finalidad de que docentes y estudiantes enriquezcan la información con etiquetas formadas de su propio conocimiento del tema.
- ✓ Se realice investigación sobre tecnologías de extracción de metadatos, orientadas a otras plataformas web 2.0 usadas en la UTPL (Wikis, CMS), para fomentar la creación de la estructura semántica de los contenidos de la UTPL.
- ✓ Fomentar el uso de estándares RDF y Dublín Core para la semantización de contenidos de las plataformas de la UTPL, con la finalidad de estandarizar este proceso con miras a desarrollar ontologías y otras tecnologías semánticas.
- ✓ Se propone como trabajos futuros que se aplique algoritmos para eliminar palabras que no son tomadas por los navegadores.
- ✓ Se propone también utilizar algún método de discriminación de caracteres como el algoritmo de la Distancia de Levenstein para controlar similitudes entre tags o sacar lexemas. En el experimento realizado para evaluar la relevancia en la extracción de tags realizado por el Sistema EMEB, se obtiene un 91,4% mediante la aplicación de F1 score en medidas de precisión, este valor es mejor cuando se combina con este tipo de algoritmos.
- ✓ Para un mejor rendimiento de la etiquetación es recomendable utilizar algoritmos de reducción de la dimensionalidad como: LSA (Latent Semantyc Analysis), Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (PCA).

GLOSARIO DE TÉRMINOS

Bookmark.- En la World Wide Web, un bookmark es una dirección Web dirección WWW o URL que queda archivada para su posterior uso, para marcar una Web interesante a fin de poder volver a él posteriormente. (ejemplo: Delicious)

Entrada.- También anotación, historia o post, designa cada una de las entradas de un weblog, puede ser un texto, una imagen, audio o vídeo. El programa de gestión del weblog le asigna de forma automática la fecha y hora de publicación, así como una dirección URL permanente o permalink.

Etiqueta (tag).- Marca que delimita un elemento en lenguaje HTML y también texto que designa el valor semántico de un objeto .Una etiqueta es una palabra clave o término (pertinente) asociado con un material informativo (como una fotografía, un artículo, un sitio Web o un vídeo clip) que describe el material. Normalmente, un elemento tendrá más de una etiqueta asociada.

Categorías.- Conjunto de secciones temáticas de un weblog, sirven como forma alternativa de navegación. Suelen estar relacionadas con la temática del blog.

HTML.– Abreviatura de HyperText Markup Language (Lenguaje de Marcación de Hipertexto). Lenguaje utilizado em la creación de páginas Web.

Plugin o plug-in.- Es un programa pequeño de computador que sirve normalmente para adicionar funciones a otros programas mayores, adicionándole alguna funcionalidad especial o muy específica.

RSS.- Abreviatura de Really Simple Syndication. Permite a los usuarios inscribirse en sitios que ofrecen "feeds" (fuentes) RSS, recibiendo información siempre que el sitio es actualizado.

Social bookmarking.- Nombre atribuido al método de almacenar, organizar, consultar y adherir favoritos de páginas Web.

XML.- EeXtensible Markup Language, es considerado un buen formato para la creación de documentos con datos organizados de forma jerárquica.

Metadatos.- datos descriptivos, tales como el autor, el título, la fecha, palabras clave, etc., asociados a un documento (o colección de documentos).

API.- APPLICATION PROGRAM INTERFACE, y representa una interfaz de comunicación entre componentes software

Blog.- (Weblog, normalmente se abrevia como blog) Blog es una abreviatura de Weblog, término utilizado para describir los sitios Web que albergan un registro constante de información.

WordPress.- Popular programa de edición de weblogs de licencia libre.

XML.- el Extensible Markup Language (XML) es un lenguaje de marcas genérico recomendado por W3C. Su objetivo básico es que sea más fácil compartir datos entre distintos sistemas de información, en particular los que están conectados a través de Internet. El XML se considera genérico porque permite a cualquiera elaborar y utilizar un lenguaje de marcas para muchos tipos de aplicaciones y dominios problema.

Taxonomía.- Los datos completos modelan en una jerarquía de la herencia donde todos los elementos de datos heredan sus comportamientos de un solo "elemento de datos estupendo". La diferencia entre un modelo de los datos y una taxonomía formal es el arreglo de los elementos de datos en una estructura arborescente formal donde está un concepto cada elemento en el árbol formalmente definido con las características asociadas.

Background.- Se dice que una aplicación funciona "en background" cuando está trabajando sin afectar a la actividad del usuario.

Dato.- Información en un formato que pueda ser procesado por un ordenador. La información se condensa digitalmente, de modo que un texto, imagen o sonido se pueda representar en la pantalla.

Estándar.- [Standard]. Norma que se utiliza como punto de partida para el desarrollo de servicios, aplicaciones, protocolos, etc.

Korn Shell.- Shell que recibió el nombre de su creador, David Korn, un investigador de Bell Laboratories de AT&T. Este shell fue presentado por primera vez en 1983 y autorizado para uso público en 1986. Se trata de una extensión de compatibilidad ascendente del shel de Bourne estándar. Todo lo que funciona con el shell de Bourne funcionará del mismo modo con Korn Shell.

RDF.- Resource Description Framework. [Marco de trabajo para la descripción de recursos]. Esquema que integra diversos metadatos, incluyendo mapas de sitios, calificación de contenido, definiciones de los canales con flujo (streaming), las colecciones de datos para los buscadores y otros conceptos, empleando la sintaxis del XML.

URL.- Universal Resource Locator. [Localizador Universal de Recursos]. Dirección de Internet que apunta a un recurso concreto dentro de un servidor conectado a la Red.

URI.- Uniform Resource Identifier. [Identificador Uniforme del recurso]. Conjunto genérico de todos los nombres y direcciones en forma de denotaciones cortas que se refieren a un recurso.

REST.- Representational State Transfer, Transferencia de Estado Representacional) – Filosofía de diseño y arquitectura web que se apoya en el intercambio de información mediante XML.

Web Semántica.- Conceptualmente se basa en añadir significado a los datos, en forma de metadatos, de modo que los ordenadores puedan entender mejor la información que existe en la World Wide Web.

Web Services.- (Servicios Web) – Conjunto de especificaciones que posibilitan la comunicación y provisión de servicios entre diferentes aplicaciones vía web.

Consorcio W3 (W3C).- Organización apadrinada por el MIT y el CERN cuyo propósito es el establecimiento de los estándares relacionados al WWW. Fue promovida por el creador del WWW, Tim Berners-Lee.

Dublin Core.- uno de los principales estándares de metadatos muy utilizado en bibliotecas y centros de documentación.

ANEXOS

ANEXO 1. ESTADÍSTICAS DE ACCESO AL WPMU-UTPL

Figura 34. Monitoreo de las suscripciones y la actividad de los elementos publicados y los elementos leídos de los blogs del WPMU-UTPL

Los días cuando se realizan más publicaciones son los siguientes:

Figura 35. Actividad de los elementos publicados y los elementos leídos según los días laborables.

ANEXO 2. SITUACIÓN DE BLOGS EN EL WPMU-UTPL

1. Estados de Entrada o (Post)

Se entiende que wordpress es un CMS que organiza sus entradas por fecha, la filosofía de Wordpress apuesta decididamente por la elegancia, la sencillez y las recomendaciones del W3C.

Dentro de las funciones del wordpress consta la de ordenar entradas y páginas estáticas en categorías, subcategorías y etiquetas. Además de proporcionar tres estados de publicación para una entrada ("post") como se muestra en la siguiente figura, cuyos estados son: **Publicado**, **Borrador y Pendiente de Revisión**.

Figura 36. Estados que puede tener una entrada dentro de un blog.

Tomando en consideración los estados de los post, se tomará como entrada válida aquellos que consten como estado "publicada".

2. Situación Actual de los Usuarios WPMU-UTPL

Wordpress maneja grupos de usuarios, los cuáles según el rol o perfil se establece distintos niveles de permisos, como se muestra en la Figura 37. Se manejan los siguientes roles:

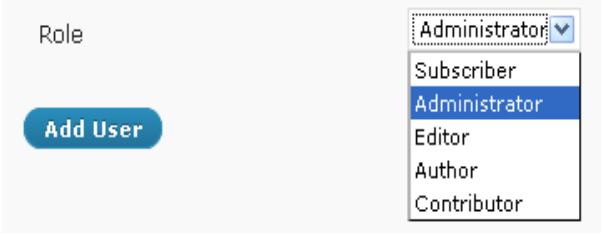


Figura 37. Roles de los usuarios del wordpress multiusuario

- ✓ **Subscriptor:** Alguien que puede leer comentarios, escribirlos, recibir boletines de noticias, entre otras cosas.
- ✓ **Administrator:** Tiene acceso a todas las características de un administrador de todo el sitio del wordpress.

- ✓ **Editor:** Puede publicar entradas, administrar sus entradas y las entradas de otros usuarios.
- ✓ **Author:** Puede publicar y administrar sus propios artículos.
- ✓ **Contributor:** Puede escribir y administrar sus propios post, sin publicarlos.

Los tipos de usuarios constan docentes, estudiantes de las modalidades que lleva la universidad, personal de los Centros Asociados de la Modalidad Abierta y a Distancia, las Asociaciones o Grupos que tiene la Universidad, así como también personal en general que desea obtener un blog.

Las entradas son almacenadas por los roles de usuario que posee el WPMU, se analizan 360 usuarios que se dividen según la Tabla 15, que al menos poseen 2 entradas en determinado blog.

Tabla 16. Usuarios del WPMU-UTPL

Usuarios	Usuario Administrador	Usuario Autor	Usuario Editor	Usuario Suscriptor
360	103	32	227	1

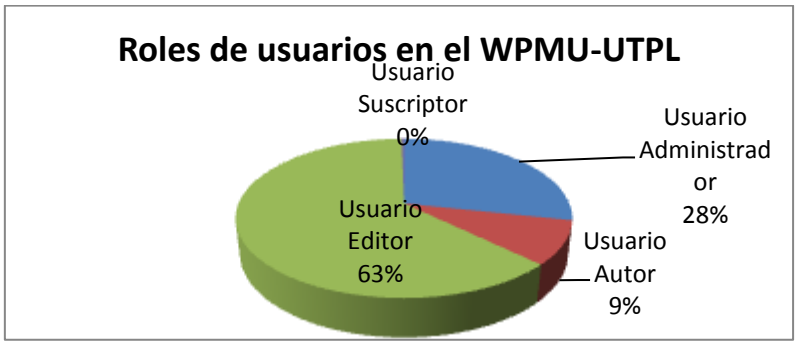


Figura 38. Usuarios del WPMU-UTPL

Existen mayoritariamente el 63% de usuarios de rol Editor, seguido por el rol Administrador con el 28% que por lo general corresponde al usuario que administra todo el blog, se visualiza apenas el 9% del rol Autor comparado con el rol Suscriptor que posee un 0%.

3. Actividad en los blogs del WPMU-UTPL

Entiéndase como ***activos*** *a aquellos blogs que publican contenido en las entradas por los usuarios* y como ***inactivos*** *a aquellos blogs que no poseen entradas por usuarios o ninguna publicación desde su creación.*

De la evaluación de los blogs se obtiene los siguientes resultados, como se representa la Tabla 15.

Tabla 17. Actividad de blogs en el WPMU_UTPL

Total	Activos	Inactivos
296	36	260

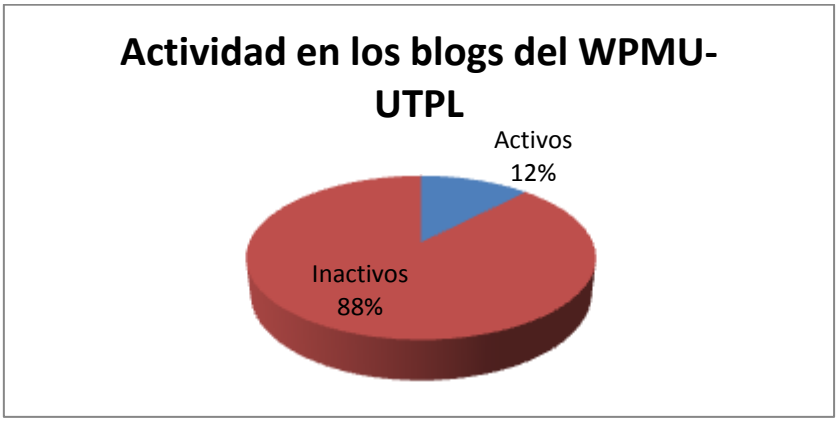


Figura 39. Actividad de los blogs del WPMU-UTPL

Tomándose los 296 blogs alojados en el WPMU-UTPL se obtiene, 12% se encuentran activos y han sido alimentados con entradas que aportan con conocimiento, en cambio se observa que el 88% de los blogs están inactivos por lo que no muestran ninguna actividad.

Es necesario enfatizar que:

- ✓ Se presenta inactividad en los blogs debido a los períodos académicos de estudio o ciclos que mantiene la UTPL, por lo que existen blogs de materias que solo se dictan en determinado ciclo académico.
- ✓ En cierta medida uno de los motivos para que existan blogs inactivos es por la creación de blogs para el proyecto "Un blog por asignatura" llevado anteriormente por el CITTES Gestión del Conocimiento, aquellos que los docentes no los utilizan por desconocimiento o por la negativa a compartir información, y no fueron dados de baja del WPMU-UTPL.

4. Etiquetación de entradas del WPMU-UTPL

La visualización de la información contenida en los blogs, se debe en gran parte al etiquetado social, donde el usuario agrega sus propias etiquetas de forma libre y no estructurada a las entradas de los blogs, facilitando la indización para que los motores de búsqueda accedan a dicha información.

Las *etiquetas o tags, o también conocidas como palabras clave relacionadas con la entrada,* en wordpress la asignación de tags *se separa por comas (etiqueta1, etiqueta2 y etiqueta3).* Ejemplo: se tiene una entrada titulada con "Introducción a la Economía" sus tags asociados hacen referencia al contenido como: *estadística, factores económicos, finanzas, banca, contabilidad, etc.* Cada una de estas etiquetas es generalmente un enlace de Internet que conduce a una página de índice que enumera todas las entradas relacionadas con esa etiqueta.

Cabe mencionar que un tag es un ejemplo de *metadato*, entiéndase como el dato que hace referencia a otro dato o a un conjunto de datos.

Para el análisis de la etiquetación de entradas se evalúan 1174 publicaciones de los 36 blogs activos

Tabla 18. Etiquetación de entradas en el WPMU-UTPL

Total de entradas evaluadas	Entradas con etiquetas	Entradas sin etiquetas
1174	443	731

Figura 40. Entradas etiquetas Vs Entradas sin etiquetas

Según el resultado del análisis representado en la Figura 40, se evidencia que existe un 38% de entradas etiquetadas, lo que significa que los usuarios que interactúan con los blogs del WPMU-UTPL, toman su tiempo en la etiquetación de cada una de las entradas lo que permite que se facilite la búsqueda de los recursos.

Por el contrario, se mantiene un porcentaje elevado en las entradas sin etiquetas con un 62% del total de la muestra, debido a que el usuario olvida colocar u omite la etiquetación, generando desorganización.

5. Categorización de entradas en el WPMU-UTPL

Para el determinar el porcentaje de uso de categorías en las entradas se analizan 1174 entradas de los 36 blogs activos, identificando el uso de categorías por entrada según la Tabla 18.

Tabla 19. Uso de categorías en entradas del WPMU-UTPL

Total de entradas	Entradas categorizadas	Entradas sin categorías
1174	810	184

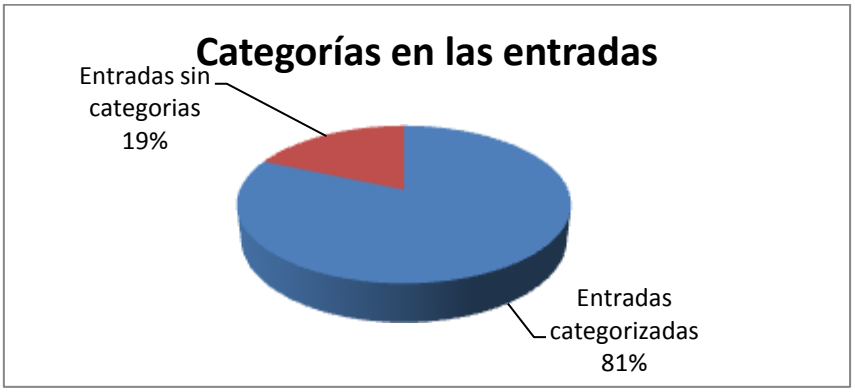


Figura 41. Categorías en las entradas del WPMU-UTPL

Según el número de entradas evaluadas y representadas en la Figura 41, existe el 81% de entradas categorizadas, lo que significa que los usuarios organizan sus contenidos al colocar categorías, y se observa que el 19% de entradas sin categorías que dificulta la visibilidad de cuyos recursos en Internet.

6. Situación Actual de los Archivos en el WPMU-UTPL

El wordpress multiusuario está configurado para permitir subir archivos con las siguientes extensiones:

Tabla 20. Tipos de archivos subidos al WPMU-UTPL

Extensión de Archivos	jpg	jpeg	png	gif	mp3	mov	avi	wmv	midi	mid	pdf	doc / docx	xls / xlsx	ppt / pptx	m3u
Imágenes	√	√	√	√											
Documentos											√	√	√	√	
Audio Video					√	√	√	√	√	√					√

Aunque se encuentra con un sinnúmero de posibilidades de archivos a ser subidos en el WPMU, no todos han sido utilizados con frecuencia. La tabla 20, presenta un resumen de los archivos publicados en entradas en el periodo Septiembre 2009 Febrero 2010.

Tabla 21. Archivos subidos al WPMU en el periodo Septiembre 2009 Febrero 2010

Blogs evaluados	Archivos	Entradas	Entradas sin archivos	Entradas con archives
36	331	1174	834	340

Figura 42. Entradas con archivos y entradas sin archivos

Las entradas que enlazan archivos del mismo dominio representan los 29%, incluidos los formatos de imágenes. En cambio el 71% son entradas sin archivos en cuyo contenido se encuentra enlaces a sitios externos código embebidos videos, presentaciones o imágenes consumidas desde de sitios externos.

Luego de el análisis de las entradas de los blogs se procede a realizar la investigación de los Sistemas de Extracción de Información a las entradas de blogs, y recoger los metadatos necesarios que identifiquen las características para semantizarlos y puedan ser reutilizados como medio de adquisición de conocimiento.

ANEXO 3. ESTUDIO DE TÉCNICAS DE DEFINICIÓN DE METADATOS COMPATIBLES CON WORDPRESS

Contenido:

1. Mediante archivos XML
2. Mediante archivos RDF
3. RSS
4. Mediante taxonomías WP TAGS SCHEMA (a nivel de tags)

1. Mediante archivos XML

XML[17] es un estándar para el intercambio de información estructurada entre diferentes plataformas. Se puede usar en bases de datos, editores de texto, hojas de cálculo, además de ser un lenguaje general de marcación que hoy en día es utilizado en diversos ámbitos sobre arquitecturas de computo, desde archivos de configuración, bases de datos, comercio electrónico y muchas más.

XML posee una tecnología sencilla que tiene a su alrededor otras que la complementan y la hacen mucho más grande y con unas posibilidades mucho mayores. Tiene un papel muy importante en la actualidad ya que permite la compatibilidad entre sistemas para compartir la información de una manera segura, fiable y fácil.

Al utilizar un XML se puede obtener lo siguiente:

- ✓ Es extensible después de diseñado y puesto en producción.
- ✓ El analizador es un componente estándar
- ✓ Al usar un documento en XML, es sencillo entender su estructura y procesarla.
- ✓ Mejora la compatibilidad entre aplicaciones, incluso permitiendo la comunicación entre aplicaciones de distintas plataformas, sin que importe el origen de los datos, es decir, podríamos tener una aplicación en Linux con una base de datos Postgres y comunicarla con otra aplicación en Windows y Base de Datos MS-SQL Server.

Sintaxis:

```
<?xml version=1.0?>
<nombre> Iliana Burguan </nombre>
<pais> Ecuador </pais>
<departamento> Sistemas </departamento>
```

2. Mediante archivos RDF

Los metadatos y **documentos XML/RDF para recuperación,** fueron aplicados en la Universidad Carlos III de Madrid [15], mediante la creación de RAI (Recuperación de Acceso a la Información) para la recuperar los archivos educativos.

Es recomendable presentar como resultado un archivo RDF el mismo que deberá servir de alimentación para futuros proyectos de acuerdo con especificaciones acorde a las

[17] Extensible Markup Language

necesidades de categorización de contenido, y es más adaptable a las tecnologías semánticas utilizadas actualmente.

La aplicabilidad de RDF (Marco de Descripción de Recursos), en la definición de recursos se debe a que es un modelo estándar para el intercambio de datos en la Web.

Un archivo RDF se basa en la idea de convertir las descripciones de los recursos en expresiones con la forma sujeto-predicado-objeto (llamadas tripletas), El sujeto es el recurso, es decir aquello que se está describiendo. El predicado es la propiedad o relación que se desea establecer acerca del recurso. Por último, el objeto es el valor de la propiedad o el otro recurso con el que se establece la relación.

Al utilizar un RDF se puede obtener lo siguiente:

- ✓ Permite denfinir recursos (cualquier cosa que pueda nombrarse mediante una URI[18]) y propiedades (característica o atributo de un recurso) y sus valores.
- ✓ Tiene asociada una URI y un significado concreto

El crecimiento de este formato es considerable ya que implícitamente es un lenguaje semántico, que brinda propiedades para compartir, colaborar y formar redes de conocimiento.

1.1 Mediante la aplicación de RDF SIOC

RDF SIOC es una tecnología de la web semántica que provee métodos para interconectar diferentes sitios de discusión de sitios web. En realidad es una ontología que se la puede reutilizar [16].

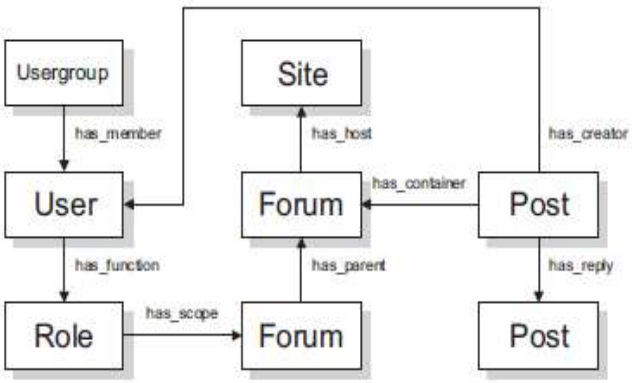


Figura 43. Ontología SIOC. [17]

Dicha ontología permite la adecuación de los documentos sobre temas específicos y sus relaciones entre: la web, foros, Post, usuarios, grupo de usuarios, y las propiedades principales que conectan entre las clases. El formato RDF es el modelo para la toma de metadatos de blogs explicita.

[18] Un URI es una cadena corta de caracteres que identifica inequívocamente un recurso (servicio, página, documento, dirección de correo electrónico, enciclopedia, etc.)

2.1.1 Experiencia con SIOC

El exportador de metadatos basados en SIOC para blogs de WordPress [18] se basa en la iniciativa que tiene como objetivo crear y potenciar una capa de datos semánticos en las comunidades en línea, basados en el formato RDF para crear descripciones de información, la misma que se publica en tuplas[19], además posee una licencia Copyright cuyos permisos para copiar y distribuir el contenido, no hay derecho para crear modificaciones o derivaciones de productos W3C con algunas condiciones de términos y condiciones.

Se implementó el plugin de SIOC en la plataforma wordpress multiusuario de la UTPL, dicho plugin consta de dos archivos los mismos que fueron copiados directamente al servidor web de la UTPL, la dirección web es http://blogs.utpl.edu.ec/.

Para apreciar su resultado en RDF, se instaló el plugin de Semantic Radar de Firefox que presenta una alarma cuando encuentra datos semánticos en sitios web, siendo necesario que el web máster habilite el uso de este plugin.

2.1.2 Activación del Plugin SIOC en Wordpress 2.7

Publicada recientemente el 25 de marzo del 2010, financiada por la Universidad Nacional de Irlanda, posee licencia Creative Commons.

Se descargan los archivos desde el sitio oficial: http://sioc-project.org/, el plugin consta de 2 archivos de en php, que son copiados en la carpeta de plugins del sitio multiusuario de la UTPL. Luego se procede a la activación para el funcionamiento del plugin.

Figura 44. Activación de SIOC en la Plataforma WPMU

Después de instalar nuevamente debido al problema de que se cierra el navegador y no se carga nuevamente desde la pag instale nuevamente

http://saeed05.wordpress.com/2009/01/11/installing-wordpress-mu-on-windows-localhost/

Y el nuevo sitio se llama multiusuario, su base de datos es multiusuariodb

[19] Las unidades de datos se denominan comúnmente tuplas

Figura 45. Activación del plugin en el Blog de Sistemas Operativos

Se realizó la activación en el blog de la materia de Sistemas Operativos cuya dirección es: http://blogs.utpl.edu.ec/sistemasoperativos/ activándole el plugin, con una publicación de dos imágenes con y sin metadatos el resultado no fue el esperado, cómo se muestra en la siguiente imagen:

Figura 46. Aplicación del plugin de SIOC, sin respuesta con Semantic Radar

Figura 47. Pruebas de SIOC cuando genera algunos datos en un formato RDF

Para la visualización de dicho RDF es necesario activar un aplicativo de Firefox llamado Semantic Radar.

Figura 48. Detección de datos SIOC mediante Semántic Radar

Cuyo resultado es el siguiente:

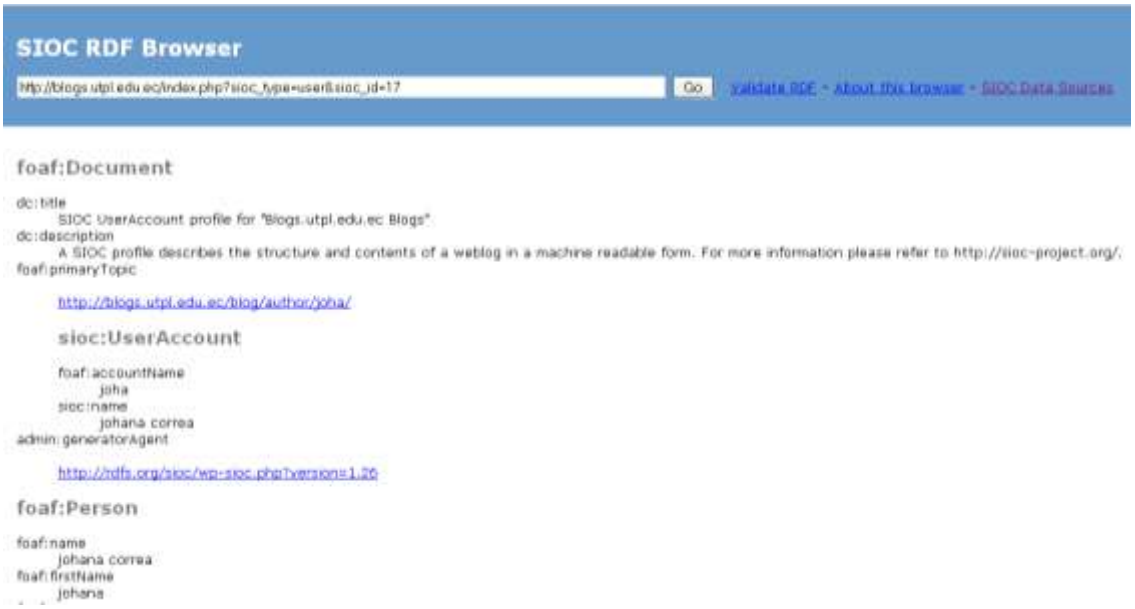


Figura 49. Resultado del plugin SIOC en el WPMU-UTPL

3. RSS

RSS[20] es una técnica para la extracción de contenido [19] La familia de RSS y tecnologías similares de suministro de noticias son el método más popular de para la distribución de entradas de blog o la obtención de metadatos sobre la estructura interna de un blog, cuya sindicación permite realizar una copia de los contenidos de un blog en otro blog (o en un lector de noticias).

El RSS es un paso más muy importante en la interconexión de la información y su acceso por los usuarios.

Es generalmente utilizado para publicar los titulares de noticias, las entradas de los blog y otras informaciones, últimamente ha creciendo en la educación mediante el desarrollo de redes hipervínculadas que abre potencialidades muy grandes en el acceso a la información.

Sin embargo, estas tecnologías están limitadas a los conceptos básicos como el título, descripción y fecha, así como un número fijo de los post publicados recientemente del blog.

4. Mediante taxonomías WP TAGS SCHEMA

WP TAGS SCHEMA es un desarrollo cuyo código puede ser utilizado en blogs mediante la modificación en la base de datos del wordpress [20].

La propuesta hace referencia al desarrollo de un sistema de taxonomías a partir de tags, basándose en un modelo de 3 tablas que están dentro del wordpress modificándoles el código manualmente, lo que permite enlazar taxonomías con los tags encontrados en una entrada del blog, y sirve como referencia para la creación de nuevos plugins que se desarrollen para WPMU.

[20] Son las siglas de RDF Site Summary or Rich Site Summary , un formato XML para sindicar o compartir contenido en la web.

4.1 Aplicación de la técnica Wp Tags Schema

La tabla wp_term contendrá todos los tags y serán almacenados de forma única sin repeticiones, y slug, que se trata de la URL que nos llevará a filtrar por ese término.

wp_terms

```
CREATE TABLE $wpdb->terms (
 term_id bigint(20) NOT NULL auto_increment,
 name varchar(55) NOT NULL default '',
 slug varchar(200) NOT NULL default '',
 term_group bigint(10) NOT NULL default 0,
 PRIMARY KEY  (term_id),
 UNIQUE KEY slug (slug)
);
```

La tabla wp_taxonomy nos sirve relacionar entre términos de categorías y tags

wp_taxonomy

```
CREATE TABLE $wpdb->terms (
 term_id bigint(20) NOT NULL auto_increment,
 name varchar(55) NOT NULL default '',
 slug varchar(200) NOT NULL default '',
 term_group bigint(10) NOT NULL default 0,
 PRIMARY KEY  (term_id),
 UNIQUE KEY slug (slug)
);
 CREATE TABLE $wpdb->term_taxonomy (
 term_taxonomy_id bigint(20) NOT NULL auto_increment,
 term_id bigint(20) NOT NULL default 0,
 taxonomy varchar(32) NOT NULL default '',
 description longtext NOT NULL,
 parent bigint(20) NOT NULL default 0,
 count bigint(20) NOT NULL default 0,
 PRIMARY KEY  (term_taxonomy_id),
 UNIQUE KEY term_id_taxonomy (term_id,taxonomy)
);
```

Es necesario registrar en una tabla la información que nos referencie una taxonomía con un objeto (post, categoría,…).

wp_term_relationship

```
CREATE TABLE $wpdb->term_relationships (
 object_id bigint(20) NOT NULL default 0,
 term_taxonomy_id bigint(20) NOT NULL default 0,
 PRIMARY KEY  (object_id,term_taxonomy_id),
 KEY term_taxonomy_id (term_taxonomy_id)
);
```

Con la modificación el código de las tablas anteriores infuye en la base de datos, se plantea un nuevo esquema de base de datos, como el esquema de taxonomía del Wordpress.

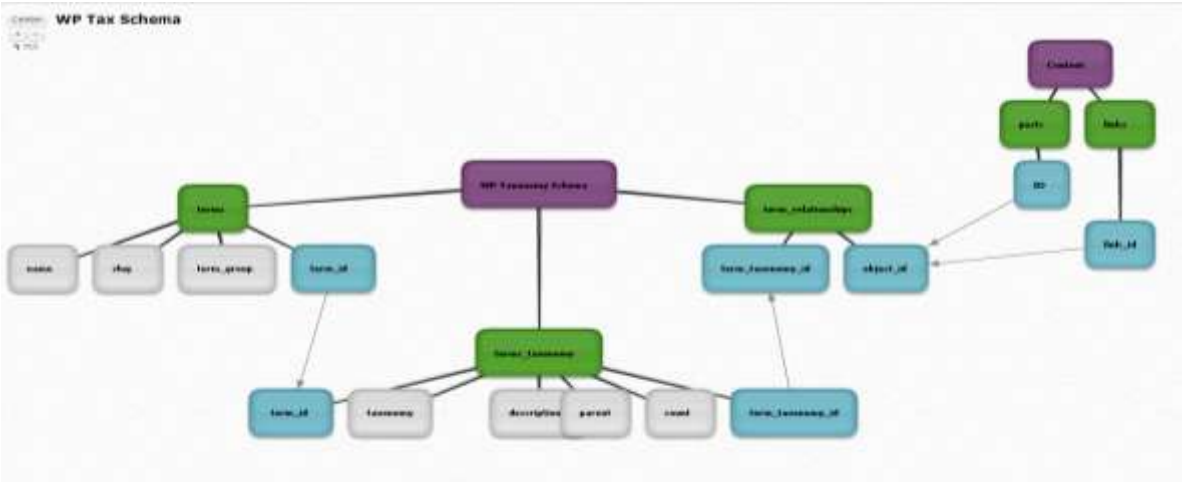


Figura 50. WP Tax SChema. [21]

Después de la evaluación de trabajo de esta técnica se evidencia que hace toma como referencia los tags ingresados por los usuarios y a su vez ordena y clasifica los tags que tengas relación con los demás posts, debido a que realiza un rastreo en todo el blogs.

ANEXO 4. ESTUDIO DE HERRAMIENTAS DE EXTRACCIÓN DE METADATOS

1. Herramientas de extracción en servidores

1.1 FOCA

Para la extracción de metadatos de archivos en la web existe la herramienta FOCA patrocinada por *Informática 641*[21] que permiten evaluar recursos relacionadas con los metadatos e información relacionada con documentos ofimáticos de Microsoft Office o Open Office.

Existe dos versiones de la herramienta FOCA, una online y una versión local aplicada a una red específica alojada en el servidor (dominio o DNS) [22].

Una prueba de Foca online analiza un archivo y devuelve como resultado la extracción de metadatos del recurso a evaluar, como se muestra en la Figura51.

Caso de prueba en Foca online de un archivo de Word

Figura 51. Extracción de metadatos de un archivo en la herramienta Foca online

[21] Empresa dedicada a actividades de formación y difusión del conocimiento técnico.

Caso de prueba en Foca en un servidor con un archivo de Word

Figura 52. Extracción de metadatos de un archivo en la herramienta Foca en un servidor

2. Herramientas de extracción de metadatos en clientes

2.1 Metadata Miner Catalogue PRO software

Este software permita extracción de metadatos de documentos de ofimática y cataloga el resumen del los archivos encontrados bajo determinada ruta, en formato HTML, pudiendo exportar a archivos RDF, CSV, SVG.

Permite la catalogación de archivo y metadatos asociados al HTML, además de la información de archivo de Adobe XMP (Extensible Metadata Plataform) de extracción de metadatos de los documentos producidos por aplicaciones de Adobe de metadatos extraídos a XML y RDF [23].

Ayuda a la visualización de las propiedades del archivo en carpetas y subcarpetas para la modificación de un conjunto de documentos.

Ventajas:

- ✓ Es una herramienta que permite automatizar la extracción de metadatos de archivos.
- ✓ Los parámetros extraídos de los archivos son: Nombre del documento, aplicación o tipo de archivo, título, autor, tema y palabras clave.
- ✓ Reúne todas las propiedades de archivos y documentos en directorios seleccionados.
- ✓ Permite la exportación en un formato XML, RDF de manera sencilla y abierta.

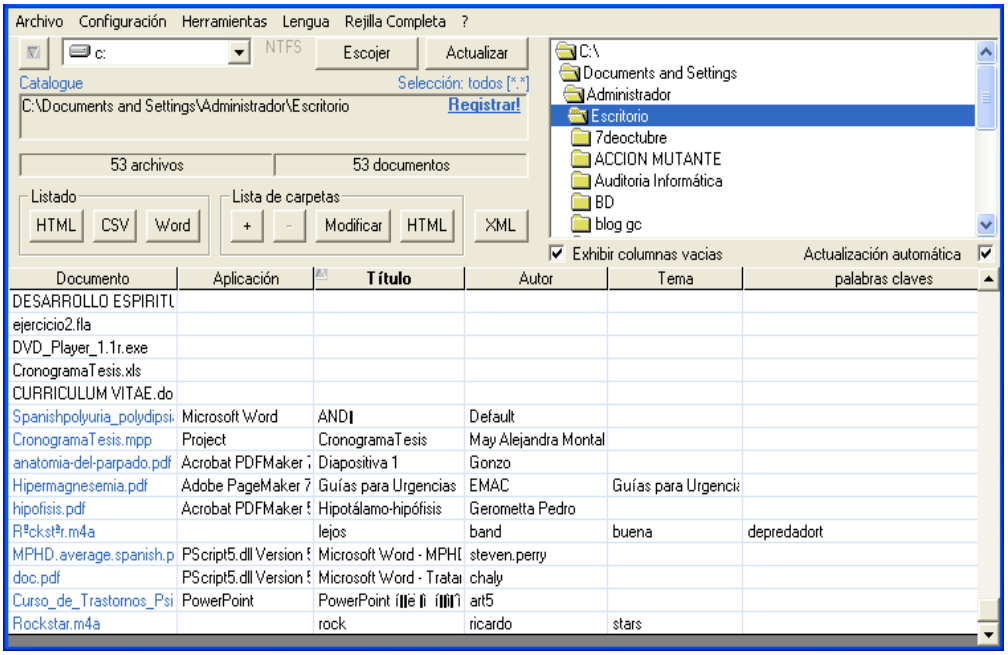


Figura 53. Prueba de Catalogue Dataminer con varios formatos de archivos y la extracción de metadatos

Además brinda la posibilidad de configuración de que metadatos a filtrar.

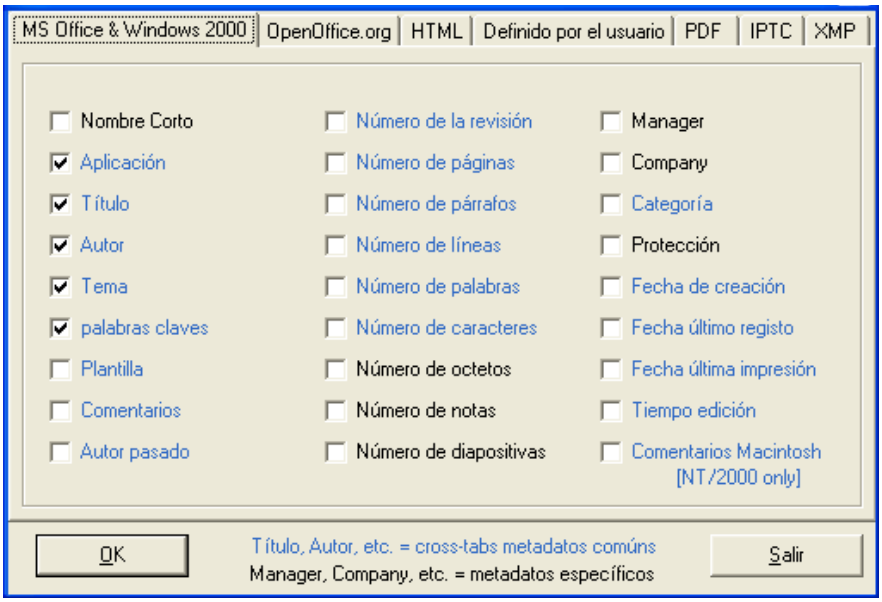


Figura 54. Configuración en Catalogue Dataminer de metadatos a presentar con varios formatos de archivos

Se puede obtener la información de resumen de MS Office, documentos Adobe y los documentos de otra fuente, y fácil de mover que los meta datos en XML o publicarlo en unidades compartidas como un catálogo de índice de la documentación HTML con enlaces de hipertexto para acceder a archivos.

C:\Documents and Settings\Administrador\Escritorio\Auditoria Informática [1/1] : 24 document(s)

1 Documento	Aplicación	Título	Autor	Tema	palab
AI_Proyecto_CursosEspecializados.doc	Office Word	Auditoria Interna - Cursos Especializados	dy3g0s		
auditoria física.ppt	PowerPoint	Sin título de diapositiva	Ronald Kleiner Toledo Macas		
AUDITORIA INFORMATICA.rar					
AUDITORIA INFORMATICARodrigo.rar					
AUDITORIA Y SEGURIDAD DE SISTEMAS 6.ppt	PowerPoint	Modelo SPICE y guía para evaluación	Jose A. Calvo-Manzano Villalón		
Auditoria.rar					
Catalogue.html		Documents catalogue			
COBIT 4 Planear y Organizar.ppt	Office PowerPoint	COBIT 4.0	Ilana Burgman		
Cobit4_Espanol.pdf	Microsoft® Office Word 2007 Versión de Evaluación	4	JAVIER DIAZ LOPEZ		
contrato%20electronico.pdf	PDFCreator Version 0.8.1	Articulo contrato electronico	DHURTADO		

Figura 55. Prueba de Catalogue de extracción de metadatos en formato HTML.

Características de Dataminer Catalogue Pro

a) Extrae metadatos de un directorio de archivos y gestiona sus metadatos.

Tabla 22. Características de Dataminer Catalogue Pro

Tipos de archivos que permite la evaluación	Metadatos que se extraen
Archivos de Microsoft Office	Nombre de la aplicación, título, autor, tema, palabras claves, plantilla, comentarios, último autor, número de revisión, número de páginas, fecha de creación, fecha del último acceso, fecha de la última impresión, tiempo de edición, obtener el tipo de archivo.
Archivos de Open Office	Título, descripción [comentarios], tema, palabras claves, creador inicial [autor], creador [autor pasado], Impreso por, fecha de creación, última fecha de guardar, imprimir Última fecha de plantilla.
Documentos HTML	(texto entre <title> y </ title>) y los metadatos de las etiquetas según el análisis de palabras clave de las páginas HTML [24].
Documentos PDF	Cifrado, versión, autor, fecha de creación, fecha modificación, productor, título, tema, palabras clave, obtener contador de páginas PDF - extracción de datos de documentos de Adobe.
Archivos de Imágenes	Nombre del objeto, Estado, Prioridad, Categoría, de consulta Categoría, palabras clave, fecha de lanzamiento, tiempo de liberación, etc.

b) Genera salidas de diferentes formatos de archivo de los catálogos predefinidos y las exportaciones de metadatos.

- ✓ Informe del HTML para metadatos de los archivos seleccionados de un directorio o un conjunto de directorios con presentación personalizada (generador de catálogo html)
- ✓ En formato XML de los informes con todos los metadatos de las carpetas y subcarpetas contenidos para compartir datos (Generar el XML del directorio de contenido de la carpeta)

- ✓ Muchas transformaciones XSL se envían con el programa - incluyendo transformaciones en RDF
- ✓ También puede rellenar bases de datos de exportación del catálogo de metadatos de archivos XML usando herramientas profesionales para la integración de datos.
- ✓ Informe de MS Word (Word 97 o Word 2000, por ejemplo) de las propiedades de metadatos seleccionado.

Como desventajas de Dataminer Catalogue se tiene:

- ✓ Es software privativo.
- ✓ La versión trial brindan 10 días de prueba,
- ✓ Cuenta con cuatro idiomas de interfaz de usuario predefinidos: Inglés, francés, alemán, portugués.
- ✓ Todos los derechos son reservados por Soft Experience
- ✓ La versión del programa ha sido creado para trabajar con Microsoft Windows 95/98/Me/2000/2003/XP/NT4.0/Vista.

Como conclusión se tiene que ésta herramienta presenta un resumen de la extracción en XML y HTML, cuyo formato puede utilizarse y reutilizar su código en el sistema de etiquetación de contenido a los blogs de la universidad, ya que las etiquetas son consideradas como metadatos.

2.2 Software HTML Code Export V1.0.0

Este software es libre y nos permite la extracción de información y contenido de un archivo determinado.

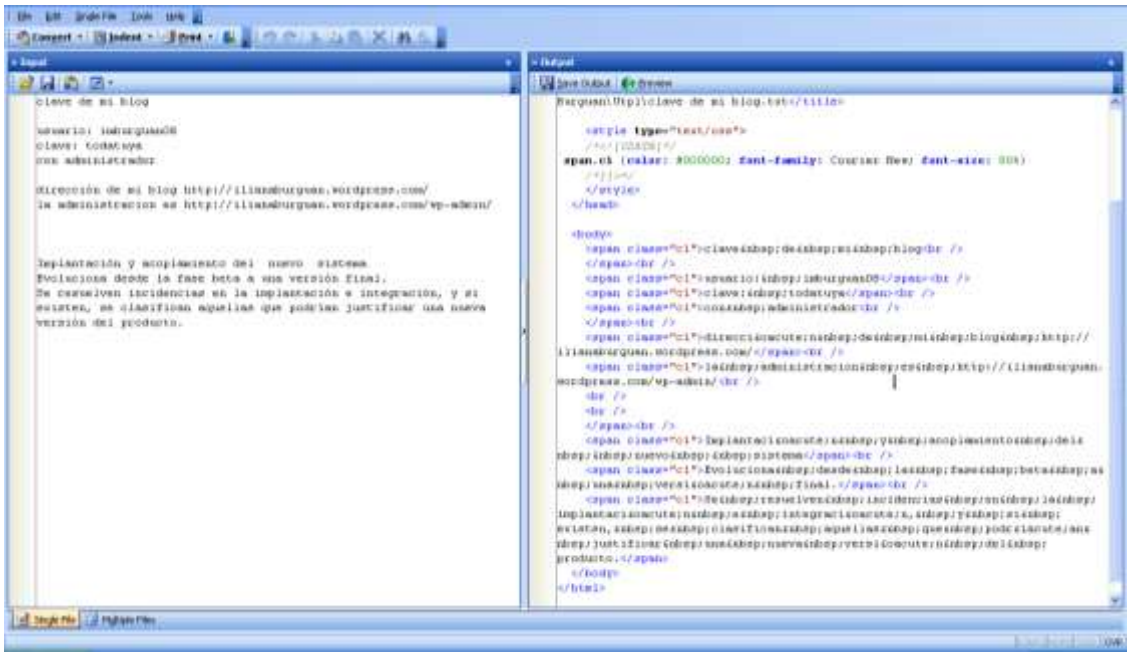


Figura 56. Extracción en Soffware HTML Code

Dentro de la ventana de configuración de la presentación de archivos de salida es configurable en: HTML, XHTML y XML.

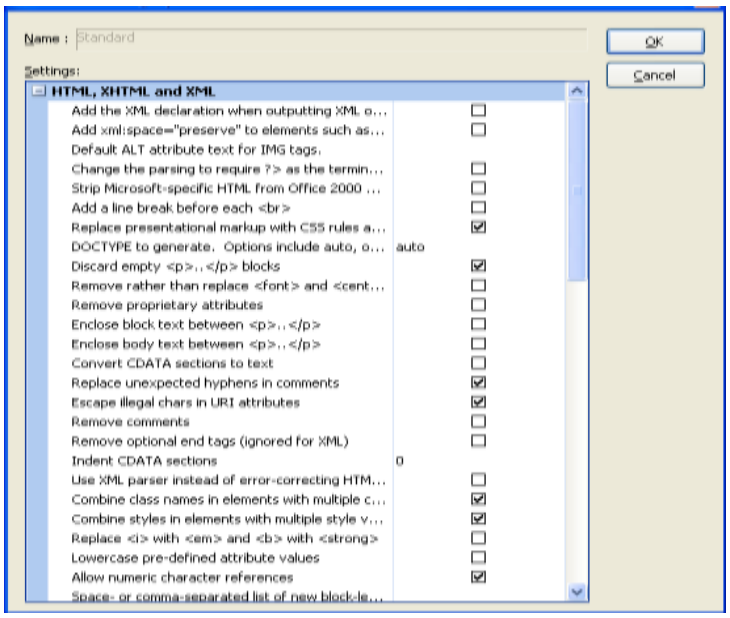


Figura 57. Configuración de formatos de examinación de archivos en Soffware HTML Code

Ventajas:

1. Convierte a HTML los siguientes formatos:

- ✓ PDF, RTF, BMP, PNG, JPG, Lotus, SVG, QUATTRO Pro, xls, xls, doc, docx, ppt, pptx y otros.
- ✓ Es 100% libre
- ✓ Crea informes avanzados mediante la opción de código de formato que son configurables en: archivos HTML, XHTML and XML, Diagnóstico, y Codificación de caracteres.

Algunos de los formatos a los que HTML Code Export convierte los ficheros HTML son los siguientes: TXT, RTF, DOC o PDF; pero también puede guardarlos en formatos gráficos como JPG, BMP o PNG.

ANEXO 5. ESTUDIO DE HERRAMIENTAS DE EXTRACIÓN DE ETIQUETAS ONLINE

1. Open Calais

Es una herramienta extrae automáticamente metadatos semánticos para contenido de sistemas de gestión de contenidos ejemplos: blogs, sitios web ya que incorpora funcionalidades semánticas dentro de un sitio web o aplicación [25].

El proceso de creación de etiquetas OpenCalais es automático teniendo un resultado en pocos segundos, usando técnicas como: procesamiento del lenguaje natural (NLP), aprendizaje de máquinas y otros métodos, Calais analiza el documento y encuentra las entidades dentro de ella.

Las etiquetas generadas se pueden incorporar a otras aplicaciones - para la búsqueda, agregación de noticias, blogs, catálogos, etc.

OpenCalais, posee un "Visor de documentos" es una forma de obtener un vistazo rápido a la salida OpenCalais; se ingresa el texto en la ventana de prueba para obtener el resultado de etiquetas. Cabe resaltar que no es un servicio web de Open Calais sino una demostración de lo que OpenCalais puede hacer.

Caso de Prueba, extracción de etiquetas

Figura 58. Caso de Prueba, extracción de etiquetas con Open Calais

Existe también un plugín basado en Open Calais para utilizarse en Wordpres, pero funciona solo en la versión de wordpress independiente cuyo nombre es Tagaro.

Calais para WordPress

El plugin Tagaroo[26] para wordpress, es una extensión de Calais para blogs. Tagaroo analiza automáticamente y propone tanto las etiquetas y las imágenes de Flickr para mejorar los post. Sugiere etiquetas, incorporarlos a tu mensaje, y realiza la búsqueda automáticamente de imágenes de Flickr para complementar cada post. Tagaroo tiene su propio sitio web. Pero actualmente no existe la versión compatible para el wordpress multiusuario de versión 2.7.

Desventajas:

- ✓ La calidad de palabras clave proporcionadas por OpenCalais varía. A veces OpenCalais no reconoce palabras claves en el texto dado.
- ✓ El clasificador OpenCalais acepta sólo los textos de tamaño limitado.
- ✓ OpenCalais clasifica correctamente sólo textos idioma Inglés.

2. AlchemyAPI

AlchemyAPI es un producto de Orchestr8, un proveedor líder de marcado semántico y soluciones de minería de texto, ayudando a las empresas a mejorar, comprender y aprovechar mejor sus activos de información-textual.

Es una herramienta que proporciona mecanismos fáciles de usar, extrae metadatos desde cualquier página web, hace la devolución de los metadatos extraídos en XML, JSON, y todos los formatos RDF, mediante los criterios de valoración de la API [27].

Para hacer uso de este recurso en un sitio web se debe descargar el ejemplo que proporciona su página y solicitar una clave de la API registrándose en su sitio oficial.

Analiza a partir de la url de un sitio web o recurso, a continuación se presenta un caso de prueba.

Figura 59. Prueba de evaluación desde una URL en el sitio oficial de Alchemy API

Así como también AlchemyAPI es capaz de extraer palabras clave de HTML, texto o contenido basado en web, cuenta con sofisticados algoritmos estadísticos y la tecnología de procesamiento del lenguaje natural para analizar sus datos, extracción de palabras clave que puede ser utilizado para indexar el contenido, y genera nubes de etiquetas.

Figura 60. Extracción de etiquetas mediante el análisis de texto en Alchemy API

ANEXO 6. REQUERIMIENTOS DEL SISTEMA EMEB

REQUERIMIENTOS FUNCIONALES:

REQ 001 Extracción de metadatos, cuando la entrada posea tags.

Entrada:

Mediante la verificación de entradas realizadas recientemente, y según la hora que se especifique para la ejecución automática, se evalúa que todos los campos estén listos y con su información completa

Proceso:

Se copian los datos que describen por cada entrada en la base de datos del wordpress multiusuario.

Salida:

Datos preparados y en espera para de la ejecución automática para pasar al siguiente proceso.

REQ 002 Extracción de metadatos, cuando la entrada no posea tags.

Entrada:

Mediante la verificación de entradas realizadas recientemente, y según la hora que se especifique para la ejecución automática, se evalúa que todos los campos estén listos y cuando el campo de tags, no posea se procede a realizar el envío de la url para la evaluación de contenido mediante el web serviceAlchemyAPI

Proceso:

Uso de la tecnología para la evaluación de lenguaje natural del recurso enviado, y la obtención de tags que devuelve dicho API. Luego se procede al almacenamiento de los datos que describen por cada entrada en la base de datos del wordpress multiusuario.

Salida:

Datos preparados y en espera para de la ejecución automática para pasar al siguiente proceso.

REQ 003Generación de Archivos RDF

Entrada:

Se inicia con la orden a partir de la ejecución automática, donde se emplea los elementos que describen los bookmarks obtenidos como title, creator, description, source, subject, identifier

Proceso:

Mediante la utilización de librerías para la generación de un archivo RDF, por cada día. En este RDF se utilizan para la descripción los algunos elementos del DublinCore, además que se pueden almacenar un número ilimitado de bookmarks, los cuales dependen de la cantidad de entradas realizadas por los usuarios en el wordpress multiusuario.

Salida:

Almacenamiento en una carpeta de archivos RDF generados, cuyos datos semantizados están listos para la representación del conocimiento en programas como scuttle, sabro.us o similares.

ANEXO 7. CASOS DE USO DEL SISTEMA EMEB

1. CU001- Selección y Adquisición de Metadatos:

Actores:

Sistema EMEB

Propósito:

Visión General: mediante la selección y adquisición de metadatos de las entradas de cada uno de los blogs del se obtiene los datos necesarios para describir los bookmarks.

Visión:

El sistema debe ser capaz de adquirir dichos metadatos para luego ser tratados.

Curso Típico de eventos:

Acción- Tarea de ejecución automática- Respuesta del Sistema EMEB

El sistema la orden de proceso a ejecutarse.

El sistema evalúa mediante el campo bTomado de la tabla wp_bookmarks que no han sido tomadas por el sistema EMEB anteriormente, además de evaluar si la entrada tiene el estado de "publicada".

Se adquiere los metadatos asociados a una entrada de un determinado blog.

Se registrará en la tabla wp_logs_bk los datos necesarios para infromar al administrador como: Título del Blog, Título de la entrada y más datos importantes.

Curso Alternativo:

En la punto 2 en caso de que no se encuentren entradas nuevas el sistema registrará en la tabla wp_logs-bk la actividad realizada.

2. CU002- Extracción o generación de etiquetas (Consumo del web Service a través del API)

Actores:

Sistema EMEB

Propósito:

Visión General: mediante la consulta del web services eran extraídos los tags y utilizados para llenar los campos que describen a los tags a entradas que no posean etiquetas.

Visión:

El sistema debe ser capaz de incorporar los tags devueltos desde el web Service Alchemy API.

Curso Típico de eventos:

Acción- Tarea de ejecución automática- Respuesta del Sistema EMEB

a) El sistema envía la URL a ser evaluada.
b) Se adquiere los tags devueltos del API

c) Se almacena en el campo bTags de etiquetas que hacen referencia a un bookmark.

Curso Alternativo:

En caso de que no se extraigan los tags por falta de palabras para la extracción de etiquetas, o en caso que no haya datos suficientes para dicho funcionamiento, se registrara en la tabla wp-logs-bk la tarea no realizada.

3. CU003- Almacenamiento en la Base de Datos

Actores:

Sistema EMEB

Propósito:

Visión General: mediante el almacenamiento de los metadatos en la tabla wp_bookmarks del Sistema EMEB que se adapta a la base de datos del WPMU_UTPL se trata de poseer los datos listos para ser descritos como bookmarks cuya información ha sido recogida del CU001: Selección y Adquisición de Metadatos, e incluso si utilizó el CU002: Extracción o generación de etiquetas en caso de que la entrada no haya poseído etiquetas.

Visión:

El sistema debe ser capaz de almacenar cada entrada de los datos obtenidos del CU001: Selección y Adquisición de Metadatos.

Curso Típico de eventos:

Acción- Tarea de ejecución automática- Respuesta del Sistema EMEB

1. El sistema almacenará inmediatamente de poseer los datos del CU001: Selección y Adquisición de Metadatos.
2. Se registrará en la tabla wp_logs_bk los datos necesarios para informar al administrador como: Título del Blog, Título de la entrada y más datos importantes.

Curso Alternativo:

En la punto 1 en caso de que no se encuentren entradas datos desde el CU003: Selección y Adquisición de Metadatos, el sistema registrará en la tabla wp_logs-bk la ausencia de datos a almacenar.

2. CU004- Generación de RDF

Actores:

Sistema EMEB

Propósito:

Visión General: mediante la generación de un archivo semántico RDF se pretende realizar la extracción de bookmarks a partir de las entradas de los blogs del WPMU_UTPL con la información categorizada y organizada que sirva como base para sistemas de representación del conocimiento de acuerdo a filtros temáticos.

Visión:

El sistema debe ser capaz de generar un archivo RDF diariamente y sea almacenado en una carpeta donde cualquier sistema de representación de bookmarks utilice ésta información contenida en archivos semánticos.

Curso Típico de eventos:

Acción- Tarea de ejecución automática- Respuesta del Sistema EMEB

1. El sistema generará un archivo semántico al finalizar el día o según como se habilite el cronograma para las ejecuciones automáticas,
2. En el archivo se describirá los datos de cada una de las entradas de datos extraídos o completados.
3. Se registrará en la tabla wp_logs_bk la generación del RDF para informar al administrador la actividad realizada.

Curso Alternativo:

En la punto 1 en caso de que no se encuentren datos en la tabla wp_bookmarks según el accionar del CU003- Almacenamiento en la Base de Datos, el sistema registrará en la tabla wp_logs-bk la ausencia de datos y por consiguiente la no generación del archivo RDF.

ANEXO 8. CARACTERÍSTICAS Y PRUEBA DE USO DEL WEB SERVICE ALCHEMY API

Contenido:

1. Llamada al API

AlchemyAPI ofrece a utilizar las instalaciones fácil para el procesamiento de su contenido: Extrae automáticamente palabras clave tema de cualquier página web o link enviado.

URLGetRankedKeywords	Se utiliza para la extracción de palabras clave de una página web de acceso público a Internet.
HTMLGetRankedKeywords	Se utiliza para la extracción de palabras clave de subido contenido HTML.
TextGetRankedKeywords	Se utiliza para la extracción de palabras clave de contenido de texto cargado.

Debido a que se el Sistema EMEB se encuentra adaptado dentro de la plataforma del wordpress multiusuario de la UTPL, y está expuesto sus publicaciones online; cualquier persona puede hacer uso de aquella información.

Se realiza la extracción de etiquetas para las entradas de los blogs que no posean palabras clave (tags), el servicio **URLGetRankedKeywords**, el mismo que utiliza una serie de parámetros para el proceso de extracción.

2. Parámetros de Alchemy para la Generación de Etiquetas

Argumento http	Descripción del parámetro
url	http url (debe ser uri-argumento codificada) (**requiere** de parámetros)
apikey	su clave privada api (**requiere** de parámetros)
maxRetrieve	número máximo de palabras clave para extraer (por defecto: 10) (Parámetro opcional)
maxNumWords	número máximo de palabras en cada palabra clave extraídos (por defecto: 3) (Parámetro opcional)
keywordExtractMode	el modo de extracción de palabras clave (normal - strict) Los valores posibles:

<table>
<tr><td></td><td>normal - palabra clave la extracción modo normal (por defecto)
strict - la palabra clave la extracción de modo estricto (devuelve más y mejores palabras clave) optimiza los resultados y retorne menos palabras clave.
(Parámetro opcional)</td></tr>
<tr><td>useMetadata</td><td>si va a utilizar palabras clave incrustadas en la página web de meta-datos. Los valores posibles:1 - enable (por defecto)0 - disable

(Parámetro opcional)</td></tr>
<tr><td>outputMode</td><td>el formato deseado de salida de la API
Los valores posibles:
xml (por defecto)
json
rdf
rel-tag
rel-tag-raw
(Parámetro opcional)</td></tr>
<tr><td>jsonp</td><td>de devolución de llamada JSONP
(Parámetro opcional, requiere un "outputMode" que se establezca en json)</td></tr>
<tr><td>showSourceText</td><td>si se incluye el texto original "de las palabras clave se extraen de la respuesta de la API.
Los valores posibles:
1 - habilitado
0 - desactivado (por defecto)
(Parámetro opcional)</td></tr>
<tr><td>sourceText</td><td>AlchemyAPI soporta múltiples modos de extracción de texto: limpieza de página web (elimina los anuncios, enlaces de navegación, etc), la extracción de texto sin formato (todos los procesos de página web, incluidos los anuncios / enlaces de navegación) y consultas XPath.
Los valores posibles:
<table>
<tr><td>cleaned_or_raw</td><td>limpieza habilitada, alternativa, al limpiar no produce ningún texto por defecto)</td></tr>
<tr><td>cleaned</td><td>limpieza habilitado , del texto de páginas web</td></tr>
<tr><td>raw</td><td>operar en texto sin formato página web (página web de limpieza disabled)</td></tr>
<tr><td>cquery</td><td>operar sobre los resultados de una disabled</td></tr>
<tr><td>xpath</td><td>operar sobre los resultados de una consulta XPath Nota: Lahttp 'XPath' también se debe establecer en una consulta XPath válida.</td></tr>
</table>
(Parámetro opcional)</td></tr>
<tr><td>cquery</td><td>Restricción de las consultas que las operaciones de la API para llevar a cabo en un área de orientación de una página web, como un título de historia o descripción del producto. (Parámetro opcional, se utiliza cuando sourceText se establece en 'cquery. uri-argument codificada)</td></tr>
<tr><td>xpath</td><td>de aplicar una consulta XPath a la página web. XPath consultas que las operaciones de la API para llevar a cabo en un área de orientación de una página web, como un título de historia o descripción del producto. (Parámetro opcional, se utiliza cuando sourceText se establece en 'xpath. Uri debe ser argumento-codificada)</td></tr>
<tr><td>baseUrl.</td><td>rel-tag de salida url http base (debe ser uri-argument codificada)
(Parámetro opcional, se utiliza con rel-tag o rel-tag or rel-tag-raw outputMode.</td></tr>
</table>

3. Formatos de Respuesta

AlchemyAPI es capaz de volver extrae tags o palabras clave de meta-datos en una variedad de formatos, incluyendo XML, JSON, microformatos, y mucho más! Más información sobre cada uno de los formatos de respuesta se proporciona a continuación:

<table>
<tr><td>XML</td><td>El formato de respuesta predeterminada.
La salida XML puede ser obligado mediante el parámetro de la API:
outputMode = xml

Formato de Respuesta (XML):

<results>
<status>REQUEST_STATUS</status>
<url>REQUESTED_URL</url>
<language>DOCUMENT_LANGUAGE</language>
<text>DOCUMENT_TEXT</text>
<keywords>
<keyword>
<text>DETECTED_KEYWORD</text>
<relevance>DETECTED_RELEVANCE</relevance>
</keyword>
</keywords>
</results></td></tr>
<tr><td>JSON</td><td>JSON hace que sea fácil de integrar la extracción de meta-datos en Javascript y tus aplicaciones web favoritas.
La salida JSON puede ser obligado mediante el parámetro de la API:
outputMode = json

Formato de Respuesta (JSON):

{
"status": "REQUEST_STATUS",
"url": "REQUESTED_URL",
"language": "DOCUMENT_LANGUAGE",
"text": "DOCUMENT_TEXT",/text>
"keywords": [
{
"text": "DETECTED_KEYWORD",
"relevance": "DETECTED_RELEVANCE"
}
]
}</td></tr>
</table>

<table>
<tr><td>RDF</td><td>RDF facilita la integración de AlchemyAPI obteniendo como salidas resultados para aplicaciones semánticas, RDF almacena tuplas y más la producción de tuplas pueden ser forzadas mediante el parámetro de la API:
outputMode = RDF
Formato de Respuesta (RDF):

<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:aapi="http://rdf.alchemyapi.com/rdf/v1/s/aapi-schema#"
xml:base="http://rdf.alchemyapi.com/rdf/v1/r/response.rdf">
<rdf:Description rdf:ID="DOCUMENT_HASH">
<rdf:type rdf:resource="http://rdf.alchemyapi.com/rdf/v1/s/aapi-schema#DocInfo"/>
<aapi:ResultStatus>REQUEST_STATUS</aapi:ResultStatus>
<aapi:URL>DOCUMENT_URL</aapi:URL>
<aapi:Language>DOCUMENT_LANGUAGE</aapi:Language>
<aapi:DocText>DOCUMENT_TEXT</aapi:DocText>
</rdf:Description>
<rdf:Description rdf:ID="DOCUMENT_HASH-KEYWORD_NUM">
<rdf:type rdf:resource="http://rdf.alchemyapi.com/rdf/v1/s/aapi-schema#KeywordOccurrences"/>
<aapi:Doc>DOCUMENT_HASH</aapi:Doc>
<aapi:Relevance>DETECTED_RELEVANCE</aapi:Relevance>
<aapi:Name>DETECTED_KEYWORD</aapi:Name>
</rdf:Description>
</rdf:RDF></td></tr>
<tr><td>Microformatos (rel-tag)</td><td>Los microformatos permiten la integración de información normalizada, estructurada en cualquier página web.
Rel-tag de salida puede ser obligado mediante el parámetro de la API:
outputMode=rel-tag
La salida de rel-tag es normalmente incrustada en una respuesta de la API en formato XML.
Para forzar el contenido Microformato rel-tag (sin una respuesta XML), utilice el parámetro de la API:
outputMode=rel-tag-raw
Formato de Respuesta (REL-TAG Microformat [XML-embedded]):

<results>
<status>REQUEST_STATUS</status>
<url>REQUESTED_URL</url>
<language>DOCUMENT_LANGUAGE</language>
<text>DOCUMENT_TEXT</text>
<microformats>
DETECTED_KEYWORD
DETECTED_KEYWORD
</microformats>
</results>

Formato de Respuesta (REL-TAG Microformat [raw]):

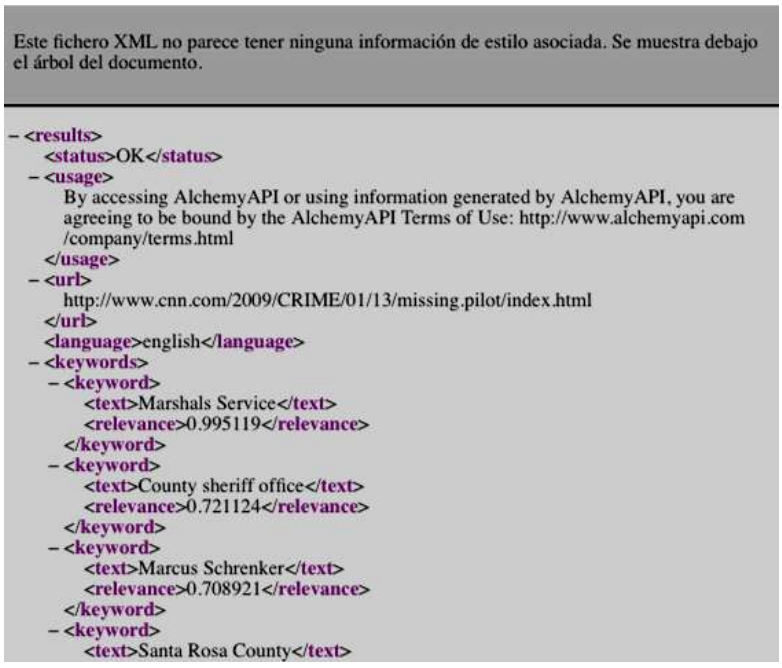
DETECTED_KEYWORD
DETECTED_KEYWORD</td></tr>
</table>

4. Respuesta de campos:

Nombre de campo	Descripción de campo
status	Éxito / fracaso del estado que indica si la solicitud ha sido procesada. Los valores posibles: **OK** **ERROR**
language	El idioma detectado que el texto original que fue escrito
url	La URL que fue solicitada
relevance	La puntuación de relevancia para una palabra clave detectados. Los valores posibles: (0,0 - 1.0) [1,0 = más pertinentes]
text	Las palabra clave detectadas en el texto.
statusInfo	Error de información de estado (enviado sólo si "el estado" == "ERROR"). Los valores posibles: **invalid-api-key** **cannot-retrieve** **page-is-not-html**

Ejemplo de Respuestas:

XML: http://access.alchemyapi.com/calls/ ...



Este fichero XML no parece tener ninguna información de estilo asociada. Se muestra debajo el árbol del documento.

```
- <results>
    <status>OK</status>
  - <usage>
      By accessing AlchemyAPI or using information generated by AlchemyAPI, you are
      agreeing to be bound by the AlchemyAPI Terms of Use: http://www.alchemyapi.com
      /company/terms.html
    </usage>
  - <url>
      http://www.cnn.com/2009/CRIME/01/13/missing.pilot/index.html
    </url>
    <language>english</language>
  - <keywords>
    - <keyword>
        <text>Marshals Service</text>
        <relevance>0.995119</relevance>
      </keyword>
    - <keyword>
        <text>County sheriff office</text>
        <relevance>0.721124</relevance>
      </keyword>
    - <keyword>
        <text>Marcus Schrenker</text>
        <relevance>0.708921</relevance>
      </keyword>
    - <keyword>
        <text>Santa Rosa County</text>
```

Figura 61. Respuesta de Alchemy API en RDF

RDF: http://access.alchemyapi.com/calls/...

Figura 62. Respuesta de Alchemy API en

5. Implementación y Prueba de Alchemy API para la extracción de Tags

Paso 1: Regístrese para obtener una clave API

Para utilizar Alchemy API, necesita una clave de API. en http://www.alchemyapi.com/api/register.html

Paso 2: Descargar un SDK

Alchemy API proporciona SDK o también llamado kit, en el lenguaje PHP (actualmente existe la versión 5.0), lo que simplifica la integración y el análisis de las capacidades de los contenidos en su aplicación.

Paso 3: Uso de Alchemy API

Un vez obtenida la clave API, se copia en un archivo de formato .txt, y todo el SDK en un servidor online o local, este kit viene con varios archivos de todo lo que proporciona para el consumo de sus servicios, pero para el caso de la extracción de tags, se utilizada el keywords.php, que está ubicada dentro de la carpeta example.

Mediante la aplicación se puede descargar de la web el ejemplo para las pruebas y adaptación del AlchemyAPI, el mismo que se lo prueba en un servidor local con la carpeta en la ruta determinada para su ejecución:

Dentro de la carpeta de ejemplo existen los siguientes directorios:

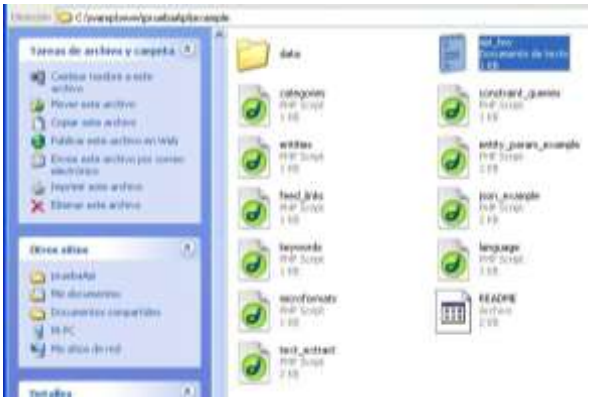


Figura 63. Directorio de carpetas para la Prueba de ALchemy API

Para realizar y evidenciar la extracción de tags mediante el Web Service de AlchemyAPI, se necesita un servidor local para poner a prueba el ejemplo para ser implementados en sistemas que se necesite evaluar una página web, blog, post, imágenes, y documentos.

Se necesita de una clave API, que se la puede obtener desde el sitio oficial de web service.

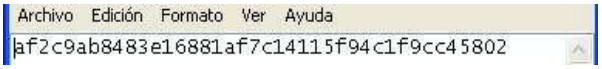


Figura 64. Clave del API

A continuación se muestra los directorios del API

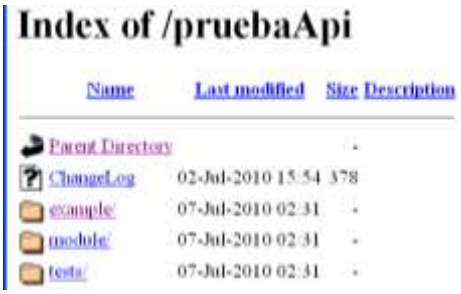
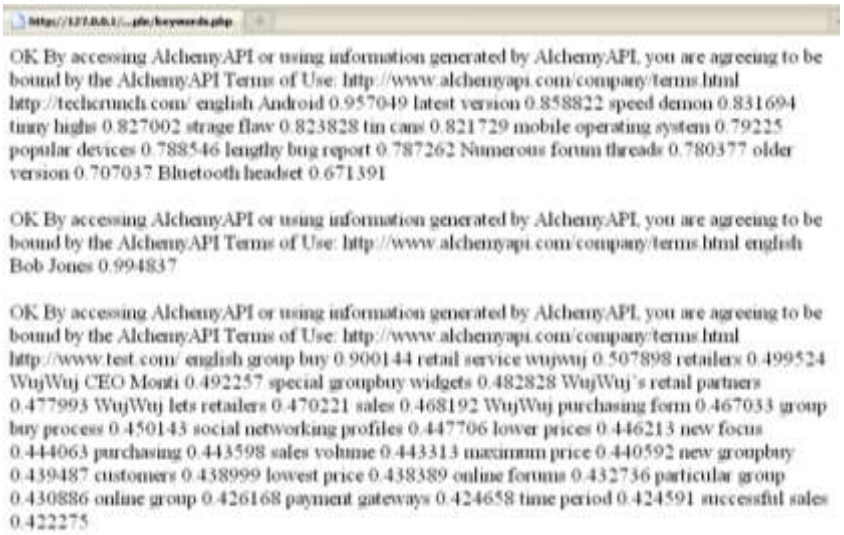


Figura 65. Directorio Pricipal de Prueba del SDK de ALchemy API

Dentro de la carpeta Example

Figura 66. Directorio de ejemplo del API

Se procede a dar click en el archivo *keywords.php* para presentar el resultado de la extracción de las etiquetas mostrada a continuación en este ejemplo está configurado con una página HTML propia para ser evaluada.



OK By accessing AlchemyAPI or using information generated by AlchemyAPI, you are agreeing to be bound by the AlchemyAPI Terms of Use: http://www.alchemyapi.com/company/terms.html http://techcrunch.com/ english Android 0.957049 latest version 0.858822 speed demon 0.831694 tinny highs 0.827002 strage flaw 0.823828 tin cans 0.821729 mobile operating system 0.79225 popular devices 0.788546 lengthy bug report 0.787262 Numerous forum threads 0.780377 older version 0.707037 Bluetooth headset 0.671391

OK By accessing AlchemyAPI or using information generated by AlchemyAPI, you are agreeing to be bound by the AlchemyAPI Terms of Use: http://www.alchemyapi.com/company/terms.html english Bob Jones 0.994837

OK By accessing AlchemyAPI or using information generated by AlchemyAPI, you are agreeing to be bound by the AlchemyAPI Terms of Use: http://www.alchemyapi.com/company/terms.html http://www.test.com/ english group buy 0.900144 retail service wujwuj 0.507898 retailers 0.499524 WujWuj CEO Monti 0.492257 special groupbuy widgets 0.482828 WujWuj's retail partners 0.477993 WujWuj lets retailers 0.470221 sales 0.468192 WujWuj purchasing form 0.467033 group buy process 0.450143 social networking profiles 0.447706 lower prices 0.446213 new focus 0.444063 purchasing 0.443598 sales volume 0.443313 maximum price 0.440592 new groupbuy 0.439487 customers 0.438999 lowest price 0.438389 online forums 0.432736 particular group 0.430886 online group 0.426168 payment gateways 0.424658 time period 0.424591 successful sales 0.422275

Figura 67. Extracción de Tag mediante Alchemy API

Extracción de etiquetas, como resultado de la evaluación del contenido, se presenta en texto por la versión del navegador, pero en realidad es un XML.

6. Extracción de tags mediante AlchemyAPI para el Sistema EMEB

Para la extracción de datos se utiliza el archivo *extraer-datos.php,* que a partir de una función que analiza el contenido de la url tomada el resultado se devuelve en código XML, el mismo que se lee y adapta para que llene los datos del campo *bTags* de la tabla *wp_bookmarks.*

La carpeta AlchemyAPI tiene una carpeta interna, dentro de la carpeta *module* se encuentra la carpeta de Alchemy Api como lo demuestra la siguiente gráfica.

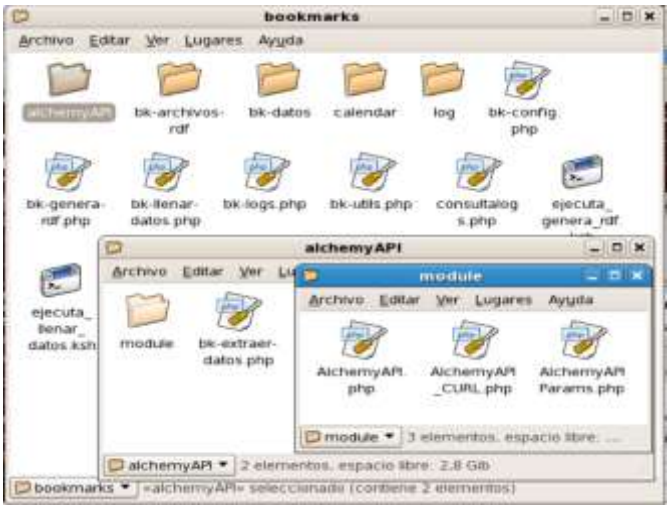


Figura 68. Directorio de archivos de Achemy API adaptados en el Sistema EMEB

7. Características a considerar del API:

- ✓ Las llamadas a URLGetRankedKeywords se puede hacer usando HTTP GET o POST.
- ✓ Las llamadas HTTP POST deben incluir el encabezado Content-Type: application / x-www-form-urlencoded
- ✓ La recuperación de URL de la cual se pide la extracción, si sobrepasa del máximo de 10 segundos, se presentará en un mensaje "no se puede recuperar".
- ✓ El envío de los documentos HTML pueden tener un máximo de 600 kilobytes. Si el documento pesa más dará lugar a un "contenido supera en tamaño-límite" y la respuesta será de error.
- ✓ La detección de idioma se realiza en el documento recuperado antes de proceder a la extracción de palabras clave. Un mínimo de 15 caracteres de texto deben existir dentro del documento solicitado HTTP para realizar detección de idioma.
- ✓ Los documentos que contengan menos de 15 caracteres de texto se supone que son los contenidos en idioma Inglés.

8. Idiomas soportados :

AlchemyAPI es capaz de extraer tags o palabras clave de los contenidos escritos en una variedad de idiomas. El apoyo multilingüe avanzado permite el etiquetado de forma automática.

La extracción de palabras clave es compatible con el contenido escrito en los siguientes idiomas: inglés, francés, alemán, italiano, portugués, ruso, español y sueco.

ANEXO 9. INSTALACIÓN, CONFIGURACIÓN Y PRUEBA DEL PLUGIN CATEGORY MAPPING

Se realizó la instalación del plugin al servidor donde se encuentra alojado la plataforma del Wordpress multiusuario de la UTPL. Se tomó a consideración que, debería tener una clasificación de blogs de diferente ámbito, ya que existe gran información y para realizar un filtrado por áreas y propósito se implemento el plugin, que sirva también para generar un filtrado y conseguir un filtrado de conocimiento.

Se agrego la clasificación a nivel superior, con las siguientes macro categorías:

1. Área Administrativa
2. Área Técnica
3. Área Socio-Humanística
4. Área Biológica
5. Blogs de Centros Asociados
6. Blogs de Departamentos
7. Blogs Personales
8. Blogs de Grupos o Asociaciones
9. Modalidad a Distancia

Pasos de la instalación y aplicación

1. Se descarga el plugin Category maping de la web
2. Se lo descomprime, se sube un nivel de carpeta y se copia la carpeta en el sitio en donde se encuentra alojado en el servidor.
3. En el panel de administración del wordpress se habilita el plugin.
4. Para añadir las categorías top o de nivel superior a los blogs se dirige a *Entradas* y escoger *Categorías*, se escribe cada una de los términos de la taxonomía.

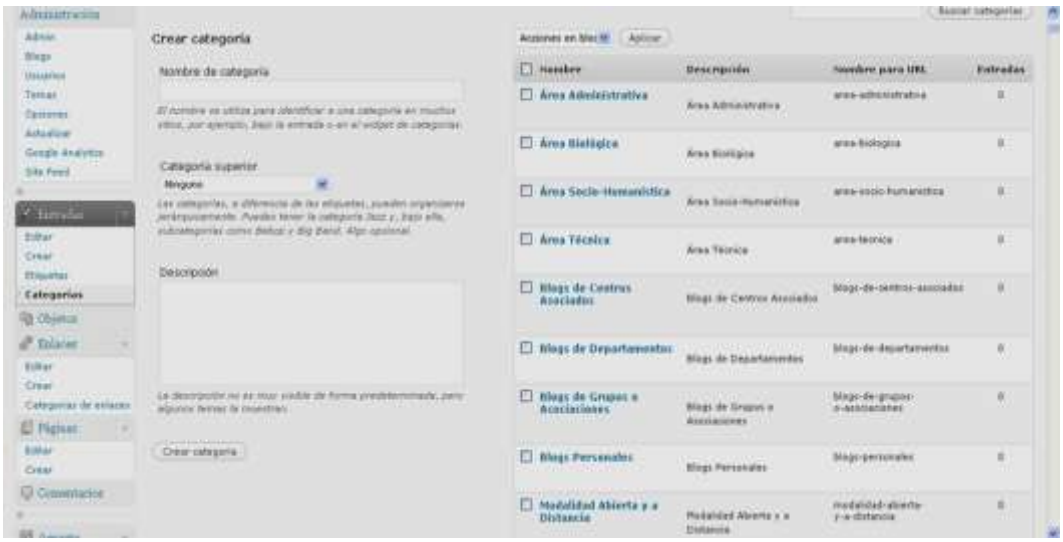


Figura 69. Administración 1 del Plugin Category Mapping

5. Luego se ingresa en la parte superior izquierda del menú del panel de administración "Plugin Category Mapping" para categorizar cada uno de los blogs según su temática.

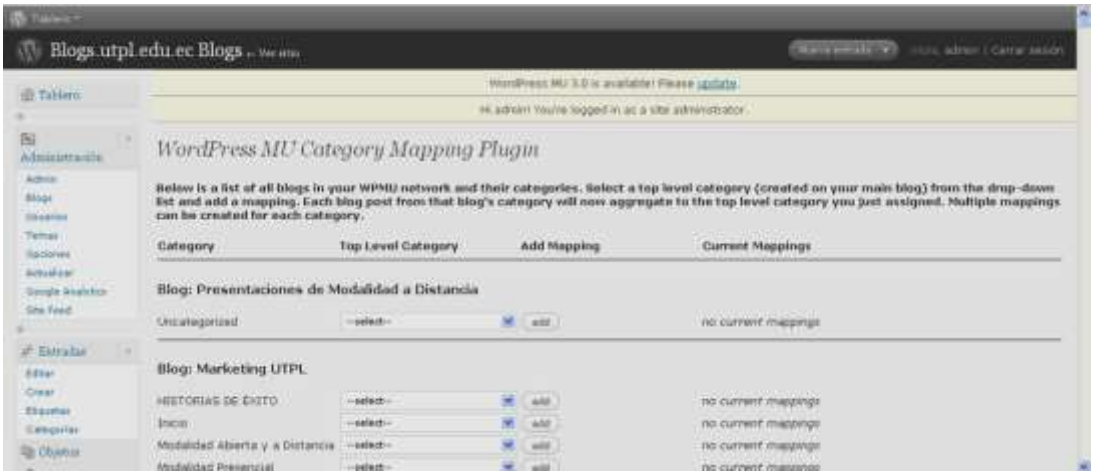


Figura 70. Administración 2 del Plugin Category Mapping

Se debe configurar el archivo del plugin alojado en la carpeta dentro del WPMU_UTPL para que permita realizar la eliminación de categorías.

Del archivo cat-plugin linea 170

```
href=/multiusuario/wp-admin/admin.php?page=cat-
plugin.php&action=del&id=".$current_map->ID."                    id=tag-check-0
class=ntdelbutton>X</a> ".$current_map->name."</span>";
```

```
164                                         AND m.sub_cat_id=".$blog_cat_ids[$key]." AND m.blog_id=".
    $blog_main_id[$key]."
165                                         ORDER BY name";
166                                     $current_maps = $wpdb->get_results($sql);
167                                     if ($current_maps) {
168                                         echo "<div id=tagchecklist>";
169                                         foreach ($current_maps as $current_map) {
170                                             echo "<span style=font-size:14px;><a
    href=/multiusuario/wp-admin/admin.php?page=cat-plugin.php&action=del&id=".$current_map->ID."
    id=tag-check-0 class=ntdelbutton>X</a> ".$current_map->name."</span>";
171                                         }
172                                         echo "</div>";
173                                     }Else{
174                                         echo "<i>no current mappings</i>";
175                                     }
176                                     ?>
177                                 </td>
178                             </tr>
179                         </form>
180                         <?php
181                         $old_blog_name = $blog_name[$key];
182                     }
183                     ?>
```

Figura 71. Modificación en el archivo del Plugin Category Mapping cat-plugin.php

ANEXO 10. CATEGORIZACIÓN DE BLOGS POR USUARIOS.

En el directorio raíz del wordpress multiusuario se modifica el archivo wp-signup.php para conseguir la opción que el usuario escoja las caterogias según estime su creador. A continuación se agrega el siguiente código después de la línea 150 dentro de la función:

function signup_another_blog ($blogname = '', $blog_title = '', $errors = '') {

```
If ($_POST["add_cat"] != "") {
            $domain=$_POST["domain"];
            $path=$_POST["path"];
            $blog_title=$_POST["blog_title"];
            $user_name=$_POST["user_name"];
            $user_email=$_POST["user_email"];
            $meta=$_POST["meta"];
            $blog=$_POST["blog"];
                $select_var = "lista_categorias";
                $select_var = $_POST[$select_var];
                $dbaux = new MySQL ();
                If ($select_var != "--seleccione--") {
                        $sql = "SELECT top_cat_id FROM wp_1_cat_mapping
                                WHERE blog_id = ".$blog."
                                AND sub_cat_id = 1
                                AND top_cat_id = ".$select_var."";
                                $chk_mapping = $dbaux->consulta ( $sql );
                        if ($dbaux->numeroFilas ( $chk_mapping ) > 0) {
                                echo "<div id=message class=updated fade>La categoria
que se desea agregar ya esta mapeada a este blog.</div>";
                        }Elseif ($select_var != ""){
                                        $sql = "INSERT INTO wp_1_cat_mapping (blog_id,
top_cat_id, sub_cat_id)
                                        VALUES (".$blog.",".$select_var.",1)";
                                        $insertar = $dbaux->consulta ( $sql );
                                echo "<div id=message class=updated fade>Categoría
agregada correctamente.</div>";
                        }
                }
                confirm_another_blog_signup($domain, $path, $blog_title, $user_name,
$user_email, $meta);
        }
        else
        {
```

ANEXO 11. ESTANDARIZACIÓN DE METADATOS

Contenido:

1. El ciclo de vida de los metadatos: creación, gestión, propagación y uso

Figura 72. Ciclo de vida de los metadatos

En la fase de creación de metadatos depende mucho de los recursos que se desee aplicar, se adapta los estándares que se acoplen al recurso, con la finalidad de que sea una norma en la adopción de metadatos en sus recursos.

En la fase de gestión, es importante planificar para el mantenimiento, incluido el control de cambio y la evaluación de impacto.

El proceso de propagación debe tener en cuenta tanto la adopción inicial del modelo de metadatos, así como actualizaciones periódicas.

3. Tipos de Estándares de Metadatos

Podemos distinguir tres tipos de estándares:

a) **Estándares de facto:** es aquel patrón o norma que se caracteriza por no haber sido consensuada ni legitimada por un organismo de estandarización al efecto.
b) **Estándares formales:** son aquellas normas formalmente establecidas por ley o por una institución reconocida para la formulación de estándares.
c) **Estándares mixtos:** es un tipo mixto de los dos anteriores.

4. Estándares de descripción de metadatos

Mediante el uso de los metadatos se trata de ubicar, dentro de Internet, los datos necesarios para describir, identificar, procesar, encontrar y recuperar un documento introducido en la red.

Los estándares más utilizados son:

- ✓ La Iniciativa de Metadatos Dublín Core (DCMI)
- ✓ LOM (Metadatos para Objetos de Aprendizaje)
- ✓ SCORM (Sharable Content Object Reference Model)

La Iniciativa de Metadatos Dublín Core (DCMI)

Describe los recursos de una manera más general, y fue desarrollado para la descripción de un amplio universo de recursos en red; su aplicación es de carácter muy general.

LOM (LearningObjectMetadata)

Es estándar especificado como 1484.12.1 IEEE *Standard for Learning Object Metadata* y especifica la sintaxis y la semántica de los metadatos de los objetos de aprendizaje. LOM, se basa en los esfuerzos previos hechos para la descripción de recursos educativos en los proyectos ARIADNE, IMS y Dublin Core. [28]

SCORM (Sharable Content Object Reference Model)

El Modelo Referenciado de Objetos de Contenido Compartible representa el conjunto de especificaciones que permiten desarrollar, empaquetar y entregar materiales educativos de alta calidad en el lugar y momento necesarios. Las especificaciones de SCORM, detallan cómo deben de publicarse los contenidos y usarse los metadatos.

4. Clasificación de los modelos de metadatos

El uso de los metadatos ha dado lugar a la aparición de la "Web semántica" la cual tiene entre sus objetivos modificar la forma en que se presenta la información de la Web de un modo que facilite su procesamiento por parte de las máquinas y de esta forma establecer canales para un factible procesamiento, integración y reutilización de la información contenida en la Web, apostando así a la extracción de conocimiento de mayor utilidad a los humanos.

Los modelos de los metadatos se clasifican en tres amplias categorías: descriptivos, estructurales y administrativos. Estas categorías no siempre tienen límites bien definidos y con frecuencia presentan un significativo nivel de superposición. Por ejemplo, los metadatos administrativos pueden incluir una amplia gama de información que podría ser considerada como metadatos descriptivos y estructurales.

Descriptivos: Usados para la descripción e identificación de recursos como Dublin Core o Etiquetas meta inmersas en el código HTML

Ejemplos:

- ✓ identificadores únicos

- ✓ atributos físicos (medios, condición de las dimensiones)
- ✓ atributos bibliográficos (título, autor/ creador, idioma, palabras claves)

Estructurales: intervienen en la recuperación de información electrónica, expresado en: **SGML[22], XML[23]/RDF[24] y EAD[25].**

Ejemplos:

- ✓ página de título
- ✓ tabla de contenidos
- ✓ capítulos
- ✓ partes
- ✓ fe de erratas
- ✓ índice
- ✓ relación con un sub-objeto (por ejemplo, fotografía de un periódico).

Administrativos: Usados en el manejo, procesamiento y administración de recursos de información en colecciones digitales. Ejemplos:

- ✓ control de acceso
- ✓ gestión de derechos
- ✓ el control de calidad y acceso
- ✓ datos técnicos sobre la creación
- ✓ utilización y condiciones de preservación

5. Normalización para el uso de Metadatos

Desde hace unos años se utiliza los modelos de metadatos para la descripción de contenidos de los documentos, que han crecido considerablemente por el auge de la web semántica de manera similar de lo que ocurre en medios impresos; se deduce que los recursos digitales son mucho más delicados ya que en la Internet cualquier persona puede ser un editor.

Actualmente existen proyectos de normas o estándares que se usan en organismos dedicados a la investigación semántica y grupos de usuarios. La ISO/IEC es la encargada de normalizar los elementos de datos y facilitar el intercambio de información entre distintas bases de datos.

---

[22] SGML son las siglas de Standard Generalized Markup Language o "Lenguaje de Marcado Generalizado". Consiste en un sistema para la organización y etiquetado de documentos. La Organización Internacional de Estándares (ISO) normalizó este lenguaje en 1986. El lenguaje SGML sirve para especificar las reglas de etiquetado de documentos y no impone en sí ningún conjunto de etiquetas en especial.

[23]Extensible Markup Language (lenguaje de marcas extensible), es un metalenguaje extensible de etiquetas desarrollado por el World Wide Web Consortium (W3C). propone como un estándar para el intercambio de información estructurada entre diferentes plataformas.

[24]El Marco de Descripción de Recursos (del inglés Resource Description Framework, RDF) es un framework para metadatos en la World Wide Web (WWW), desarrollado por el World Wide Web Consortium (W3C).

[25] Estándar para codificar instrumentos de descripción archivística por medio de SGML y XML, mantenido en la Network Development and MARC Standards Office de la Library of Congress en colaboración con la Society of American Archivists (EAD Working Group).

Por tanto hay diferentes propuestas al respecto, como sectores y organismos internacionales se encuentran desarrollando diversos esquemas y estándares para formalizar el uso de metadatos y establecer un conjunto de reglas semánticas, sintácticas y de contenido que pretenden describir homogéneamente Objetos de Aprendizaje[26] o recursos de información [29].

Los diferentes tipos de materiales digitales, y los diferentes sistemas de archivos, se necesita un soporte de metadatos diferentes, y se debe tomar en cuenta que al realizar grandes proyectos de repositorios se puede tornar caótica la organización debido a la integración de bases de datos, que tienen diferentes elementos como descripción de sus recursos en línea; al ser poco o nada entendible como se describen los recursos, es por eso que se debe utilizar un estándar con la finalidad que cualquier sistema de la misma índole, sea escalable y mejore el ámbito de cooperación entre comunidades que buscan el bien común. Ejemplo *la educación.*

[26]Un objeto de aprendizaje es cualquier agrupación de materiales que está estructurada de una forma significativa y está asociado a un objetivo educativo. los objetos de aprendizaje a los que se refiere aquí, corresponden a objetos digitales que pueden ser distribuidos mediante el blog multiusuario de la UTPL

ANEXO 12. APLICACIÓN PRÁCTICA DEL FORMATO DUBLIN CORE EN UNIVERSIDADES Y CENTROS DE INVESTIGACIÓN

Debido al crecimiento de aplicaciones con Dublín Core por parte de organizaciones dedicadas a la educación como universidades, centros de investigación y bibliotecas, y en vista que los elementos que describen recursos que sirven de conexión entre diferentes proyectos, aumentando la red vinculada a la explotación de conocimiento.

El repositorio de documentos de artículos y revisas de España denominado RedIRIS, hace uso de los meditados para la normalización de recuperación de información con DCMI patrocinada por SEDIC (Sociedad Española de Documentación e Información Científica).

La RedIRIS[27], que cuanta con más de 350 instituciones afiliadas, principalmente universidades y centros públicos de investigación. [30]

La aplicación práctica del formato Dublin Core se ha proyectado en varias grandes bibliotecas como la biblioteca del Congreso de Washighton, la Biblioteca Nacional de Australia y la Biblioteca Nacional de Nueva Zelanda, que han tratado de integrar los elementos del Dublin Core en sus catálogos.

Existen varios software creadores o generadores de metadatos como : el modelo el DC-dot o generador Dublin Core, es un creador automático Dublin Core del UKOLN [31] desarrollado por British Library Research and Innovation Centre con colaboración de la Universidad de Baht. Desarrollado en Perl y Java.

La Biblioteca Nacional de Suecia ha llegado a un acuerdo a desarrollar nuevas soluciones técnicas, con el fin de gestionar esta circulación de archivos de texto y metadatos y ofrecen en diferentes formatos: MARC21, ONIX, Dublín Core, etc. [32]

La universidad de Cambrigde, Inglaterra, usa 4 elementos Dublin Core en sus recursos educativos. Como meta tags básicos que son tratados como sinónimos por la Universidad como una guia del sitio:

- ✓ date & dc.date
- ✓ publisher & dc.publisher
- ✓ description & dc.description
- ✓ keywords & dc.subject[33]

La Biblioteca de la Universidad de Chicago, utiliza el estándar Dublin Core con algunos elementos como Dc.Creador DC.Contributor y DC.Date se utilizan para describir el material de sus recursos educativos.[34]

Los metadatos son usados en repositorios publicados en el campo de la biología, como es el caso de "Dryad" de la universidad de Carolina del Norte en Chapel Hill. EEUU., El análisis se basa en la herencia de metadatos adoptados en el campo de la biología cuyos resultados de los informes de un experimento en términos de cartografía incluye 12 vocabularios y 600 términos de aproximadamente.[35].

[27] Es la red académica y de investigación española y proporciona servicios avanzados de comunicaciones a la comunidad científica y universitaria nacional de España

ANEXO 13: TABLAS DEL SISTEMA EMEB A ADAPTAR AL WPMU-UTPL

Para que la Base de Datos no se alterada y se acople a la forma de organización del wordpress multiusuario, es necesario adaptar dos tablas mediante un sript sql, las cuales constan de los campos necesarios para la adquisición de datos necesarios para la etiquetación de bookmarks.

El Sistema EMEB posee dos tablas que se añadirán al WPMU-UTPL, las cuáles son:

- ✓ wp_bookmarks
- ✓ wp_logs_bk

Tabla wp_bookmarks

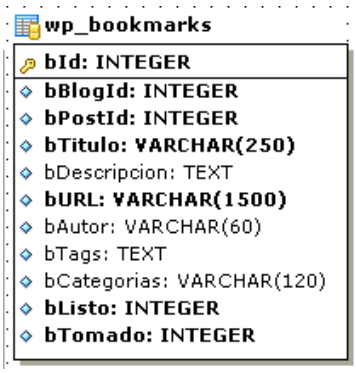


Figura 73. Tabla wp_bookmarks del Sistema EMEB

La tabla agregada se llama WP_BOOKMARKS y contendrá todos los datos procesados de los bookmarks a ser exportados al Scuttle, contiene los siguientes campos:

- ✓ **bId:** Id secuencial del bookmark
- ✓ **bBlogId:** Id del Blog al que pertenece el bookmark, como se trata de Wordpress multiusuario, existen muchos blogs, cada blog tiene un Id diferente.
- ✓ **bPostId:** Id del Post en el Blog al que pertenece el Bookmark
- ✓ **bTitulo:** Título del Bookmark.
- ✓ **bDescripcion:** Descripción del bookmark.
- ✓ **bURL:** Dirección web del bookmark
- ✓ **bAutor:** Autor del Bookmark
- ✓ **bTags:** Etiquetas del bookmark, separadas por comas
- ✓ **bCategoría:** Categorías a la que pertenece el bookmark, separadas por coma
- ✓ **bListo:** Indica si el bookmark ya contiene toda la información, o aún está incompleto.0=Información incompleta; 1=Información Completa.
- ✓ **bTomado:** Campo para la verificación que indica si el bookmark fue exportado a un archivo RDF o no'.

También se han agregado índices a la tabla para agilizar el proceso de búsqueda (índice al campo bBlogId).

Para agregar esta tabla, se debe ejecutar el siguiente script en la bd del WordPress:

```
CREATE TABLE `wp_bookmarks` (
  `bId` int(11) NOT NULL AUTO_INCREMENT COMMENT 'Id del bookmark',
  `bBlogId` int(11) NOT NULL COMMENT 'Id del blog al quepertenece el bookmark',
  `bPostId` int(11) NOT NULL COMMENT 'Id del Post en el Blog al que pertenece el Bookmark',
  `bTitulo` varchar(250) NOT NULL COMMENT 'TÃ-tulo del bookmark',
  `bDescripcion` text COMMENT 'Descripción del bookmark',
  `bURL` varchar(1500) NOT NULL COMMENT 'URL del bookmark',
  `bAutor` varchar(60) DEFAULT NULL COMMENT 'Autor del Bookmark',
  `bTags` text DEFAULT NULL COMMENT 'Tags del bookmark separado por comas',
  `bCategorias` varchar(120) DEFAULT NULL COMMENT 'Categoría del bookmark',
  `bListo` int(1) NOT NULL DEFAULT '0' COMMENT '0=Información incompleta; 1=Información Completa',
`bTomado` int(11) NOT NULL DEFAULT '0' COMMENT 'Indica si el bookmark fue exportado a un archivo RDF o no',
PRIMARY KEY (`bId`),
  KEY `i_bBlogId` (`bBlogId`)
) ENGINE=MyISAM AUTO INCREMENT=1 DEFAULT CHARSET=utf8 COMMENT='Datos procesados de los post para ponerlos como bookmarks';
```

Tabla wp_logs_bk:

Figura 74. Tabla wp_logs_bk del Sistema EMEB

La segunda tabla WP_LOGS-BK se utiliza para que el administrador pueda obtener información de los logs que registran que tipo de proceso ha realizado el servidor en cuanto a las ejecuciones automáticas, como es el caso de llenar datos y generar a RDF, y se pueda visualizar en una interfaz los registros realizados así como también el blog y post utilizado.

- ✓ **Lid:** Id secuencial del log
- ✓ **LTipo:** Se especifica 4 tipos de logs como: error, belong, info y warning
- ✓ **LDescripcion:** Se describirá acerca del tipo de ejecución hecha.
- ✓ **LPagina:** Tipo de ejecución de los archivos .php que se ejecut.
- ✓ **IblogId:** Nombre del Blog desde donde se lee la entrada (el nombre esta como ID)
- ✓ **IPostID:** Nombre de la entrada que se lee (el nombre esta como ID)

Se usa la herramienta SQL Manager 2010 para la visualización y administración de base de datos, a través de este software ejecutamos el script cuya respuesta ha sido exitosa.

Para agregar esta tabla, se debe ejecutar el siguiente script en la bd del WordPress:

```
CREATE TABLE `wp_logs_bk` (
  `lId` int(11) NOT NULL AUTO_INCREMENT COMMENT 'Identificador de Log',
  `lTipo` varchar(10) NOT NULL COMMENT 'Tipo de log: info, error, debug, warning',
  `lDescripcion` text COMMENT 'DescripciÓn del Log',
  `lPagina` varchar(100) DEFAULT NULL COMMENT 'PÃ¡gina del log',
  `lFecha` datetime NOT NULL COMMENT 'Fecha de log',
  `lBlogId` int(11) DEFAULT NULL COMMENT 'Id de Blog',
  `lPostId` int(11) DEFAULT NULL COMMENT 'Id del Post',
  PRIMARY KEY (`lId`),
  UNIQUE KEY `lId` (`lId`)
) ENGINE=MyISAM DEFAULT CHARSET=utf8 COMMENT='Tabla en donde se almacenan los logs del mÓdulo Bookmarks';
```

1. Para llenar esta tabla realizamos lo siguiente:

1. Obtener los IDS de todos los blogs que se encuentran creados y activos en el WordPress Multiusuario:

```
select blog_id
from `wp_blogs`
where delete d=0 and blog_id>1;
```

2. Por cada Id obtenido se realizar la siguiente consulta:

```
select  p.id, p.post_title, u.display_name, p.guid
from wp_2_posts p, `wp_users` u
where p.post_type='post' and p.post_author=u.ID and p.id>1
```

Se cambia el id del blog por el nombre de la tabla. Ejemplo: la tabla:

wp_2_posts indica que pertenece al blog con id=2.

En la consulta se obtene los siguientes datos:

- ✓ Id del post en el blog al que pertenece el bookmark.
- ✓ Título del Post
- ✓ Autor del Post
- ✓ URL del post

ANEXO 14. RESTRICCIONES PARA EL LLENADO DE DATOS DEL SISTEMA EMEB

Contenido:

1. Configuración del Código de Caracteres
2. Para la limpieza de etiquetas HTML utilizadas para llenar la descripción de un bookmark
3. Para cada blog obtenemos el máximo id del Post que tenemos en la base de datos
4. Consultar todos los posts de Blog, y que anteriormente no hayan sido subidos

1. Configuración del Código de Caracteres

Debido al idioma por defecto que está configurado el API, es necesario que se controle y cambie de codificación a utf-8, para que funcione de manera esperada con tildes y ñ. Este archivo se lo encuentra en el bk-util-php.

```
//dada un string determina si está en utf-8 o no
function is_utf8($Str) {
 for ($i=0; $i<strlen($Str); $i++) {
  if (ord($Str[$i]) < 0x80) continue; # 0bbbbbbb
  elseif ((ord($Str[$i]) & 0xE0) == 0xC0) $n=1; # 110bbbbb
  elseif ((ord($Str[$i]) & 0xF0) == 0xE0) $n=2; # 1110bbbb
  elseif ((ord($Str[$i]) & 0xF8) == 0xF0) $n=3; # 11110bbb
  elseif ((ord($Str[$i]) & 0xFC) == 0xF8) $n=4; # 111110bb
  elseif ((ord($Str[$i]) & 0xFE) == 0xFC) $n=5; # 1111110b
  else return false; # Does not match any model
  for ($j=0; $j<$n; $j++) { # n bytes matching 10bbbbbb follow ?
   if ((++$i == strlen($Str)) || ((ord($Str[$i]) & 0xC0) != 0x80))
   return false;
  }
 }
 return true;
}
// FUNCIÓN DE UTF8 A ISO
function UTF8toISO($string){
       if(!is_utf8($string)){
            return $string;
       }else{
            return utf8_decode($string);
       }
}
// FUNCIÓN DE ISO A UTF8
```

```
function ISOtoUTF8($string){
      if(is_utf8($string)){
          return $string;
      }else{
          return utf8_encode($string);
      }
}
```

2. Para la limpieza de etiquetas HTML utilizadas para llenar la descripción de un bookmark

Para la llenado de datos en el campo de descripción es necesario que la información se presente lo más limpia posible, es decir, sin etiquetas HTML que se aprecian sin estética visual.

A continuación se presenta el script que realiza la limpieza de la descripción de un bookmark.

```
function limpiarTexto($texto) {
        //quita las etiquetas html
        $texto = trim ( strip_tags($texto ));
        //quita el código de los objetos embebidos dentro del post
        $textoLimpio = false;
        while ( ! $textoLimpio ) {
                $cInicio = strpos ( $texto, '[caption' );
                $cFin = strpos ( $texto, 'caption]' );
                if ($cInicio !== false && $cFin !== false) {
                        if ($cInicio < $cFin + 8) {
                                $texto = str replace ( substr ( $texto, $cInicio
$cFin + 8 ), "", $texto );
                        } else {
                                $textoLimpio = true;
                        }
                } else {
                        $textoLimpio = true;
                }
        }
        //retorna el texto limpio
        return $texto;
}
```

3. Para cada blog obtenemos el máximo id del Post que tenemos en la base de datos

```
select IFNULL (MAX(bPostId),1) maxPostId from wp_bookmarks where bBlogId=
```

```
        $bkIdBlog";
```

4. Consultar todos los posts de Blog, y que anteriormente no hayan sido subidos

Según el análisis de los campos de las tablas wp_#_post y wp_user, se realizan la siguiente consulta para la extracción que sirven de datos de llenado de la tabla wp_bookmarks.

Mediante el campo blog_id de la tabla wp_blogs se filtra los blogs y se accede a sus post.

```
$strConsultaPosts = "select  p.id id, p.post_title titulo, u.display_name autor, p.guid
URL, p.post_content descripcion  from wp_" . $bkIdBlog . "_posts p, `wp_users` u  where
p.post_type='post'    and    p.post_status='publish'    and    p.post_author=u.ID    and
p.id>$maxPostId order by p.id";
```

ANEXO 15. CONFIGURACIÓN DEL SISTEMA EMEB

1. Ruta de almacenamiento de archivos RDF
2. Ejecución automática de los procesos:
3. Consulta de logs

Los archivos necesarios para el llenado de datos y la generación del RDF están inmersos dentro de una carpeta, la misma que se añade a la raíz del sitio del wordpress multiusuario llamada "bookmarks" en la que contiene varios archivos programados en php y dos adicionales de formato ksh.

1. RUTA DE ALMACENAMIENTO DE ARCHIVOS RDF

El almacenamiento de los arvichos RDF del Sistema EMEB, se almacena en un directorio bookmarks dentro del servidor donde está alojado el Wordpress multiusuario. El nombre de la carpeta es bk-archivos-rdf, que puede ser llamada desde sistemas de representación de bookmarks ejemplo: Semantic Bookmarks UTPL[28], o por repositorios similares.

```
/** Path que indica el lugar en dónde se ubicarán los archivos rdf que se
creen */

define('RDF_PATH_ARCHIVOS', '/opt/lampp/htdocs/wordpress/bookmarks/bk-
archivos-rdf/');
```

Los archivos RDF se almacenan con el nombre *"bookmarks_"* seguido de la fecha de ejecución ejemplo:

- ✓ bookmarks_20100809, *lo que significa que se registro el año 2010, el octavo mes en el día 9*. Como se muestra en la siguiente figura.

2. EJECUCIÓN AUTOMÁTICA DE LOS PROCESOS

Debido a que el servidor de pruebas donde se implementará el Wordpress Multiusuario no tiene instalado el CRONTAB[29], se han creado dos scripts ksh (Korn Shell) que permiten ejecutar los procesos de forma diaria y calendarizada; estos scripts son:

1. ejecuta_llenar_datos.ksh→ Ejecuta el proceso bk-llenar-datos.php
2. ejecuta_genera_rdf.ksh→ Ejecuta el proceso bk-genera-rdf.php

Script ejecuta_llenar_datos.ksh

[28] Semantic Bookmarks UTPL, Tesis (Lorena Leon,2010)

[29] Cron es un demonio que ejecuta tareas de manera programada basado en la configuración del comando crontab

Este script permite la ejecución automática del proceso **bk-llenar-datos.php**, cada vez que ejecuta este proceso escribe en un archivo de log dentro del directorio /bookmarks/log, en estos logs se indicará la fecha y hora en que se ejecutó. Ejm:

```
Ejecutando archivo
Fecha Inicio: 2010/08/13 23:40:01
  Total bookmarks nuevos encontrados: 9
Fecha Fin    : 2010/08/13 23:40:36
```

Se generará un archivo de log por mes, en el siguiente formato:

LogLlenarDatos_[aaaamm].log

Ejemplo: Log generado para el año 2010, mes Agosto: LogLlenarDatos_201008 log

Dentro del script existen algunos parámetros, los cuales hay que configurarlos de acuerdo a nuestras necesidades y a datos del servidor de producción.

```
############################
##### CONFIGURACION  ######
############################

# Ruta en donde se encuentra el ejecutable de php
RUTA_PHP="/opt/lampp/bin/php"

#Horas en las que se va a ejecutar el proceso (Formato de dos digitos:
Rango:  00 01 02 .. 22 23)
#Si se desea ejecutar el proceso en varias horas, escribir las horas
separadas por espacios
HORAS_EJECUCION='03 09 12'

#Minuto en la que se va a ejecutar el proceso (Formato de dos digitos:
Rango:  00 01 02 .. 58 59)
MIN_EJECUCION='00'

# Prefijo del archivo de log
PREFIJO_LOG="log/LogLlenarDatos"
```

Para los parámetros que se tiene configurado, indica que el proceso bk-llenar-datos.php se va a ejecutar todos los días a las 03H00, 09H00 y 12H00.

Una vez que se tiene correctamente configurado el script, se procede a ejecutarlo desde la consola de Linux por primera vez en background, de la siguiente manera:

1. Ubicarse en el path en donde se encuentra el script:
   cd wordpressmu/bookmarks/

2. Ejecutar el script en background:
   ksh nohup ejecuta_llenar_datos.ksh&

3. Validar de que el script se encuentre corriendo:

ps -ef | grep llenar | grep -v grep

El resultado de ejecutar este último comando debe ser similar al siguiente:
root 11542 10821 0 23:10 pts/1 00:00:00 ksh ejecuta_llenar_datos.ksh

En donde **11542** es el Id del proceso en Linux que se está ejecutando correctamente en forma silenciosa.

Si por algún motivo se necesita cambiar algún parámetro de configuración del script, lo que se debe hacer es matar el proceso actual de Linux, cambiar la configuración y nuevamente ejecutar los tres pasos anteriores.

Para matar el proceso actual ejecutamos lo siguiente:

1. Obtenemos el Id del proceso actual:
   ps -ef | grep llenar | grep -v grep

   Debe arrojarnos un resultado similar al siguiente:
   root **11542** 10821 0 23:10 pts/1 00:00:00 kshejecuta_llenar_datos.ksh
2. Matamos el proceso:
   kill -9 **11542**

Script ejecuta_genera_rdf.ksh

Este script permite la ejecución automática del proceso **bk-genera-rdf.php**, cada vez que ejecuta este proceso escribe en un archivo de log dentro del directorio /bookmarks/log, en estos logs se indicará la fecha y hora en que se ejecutó. Ejm:

```
Ejecutando archivo
Fecha Inicio: 2010/08/13 23:22:01
  Total de bookmarks tomados para generar archivo RDF:  3
Fecha Fin    : 2010/08/13 23:22:04
```

Se generará un archivo de log por mes, en el siguiente formato: LogGeneraRdf_[aaaamm].log

Ejemplo:

Log generado para el año 2010, mes Agosto: LogGeneraRdf _201008 log

Dentro del script existen algunos parámetros, los cuales hay que configurarlos de acuerdo a nuestras necesidades y a datos del servidor de producción.

```
############################
#####  CONFIGURACION  ######
############################

# Ruta en donde se encuentra el ejecutable de php
RUTA_PHP="/opt/lampp/bin/php"
```

```
#Horas en las que se va a ejecutar el proceso (Formato de dos dígitos:
Rango:  00 01 02 .. 22 23)
#Si se desea ejecutar el proceso en varias horas, escribir las horas
separadas por espacios
HORAS_EJECUCION='03 09 12'

#Minuto en la que se va a ejecutar el proceso (Formato de dos digitos:
Rango:  00 01 02 .. 58 59)
MIN_EJECUCION='20'

# Prefijo del archivo de log
PREFIJO_LOG="log/LogGeneraRdf"
```

Para los parámetros que se tiene configurado, indica que el proceso bk-genera-rdf.php se va a ejecutar todos los días a las 03H20, 09H20 y 12H20.

Se recomienda que se configure este script (ejecuta_genera_rdf.ksh) de tal manera que la hora de ejecución sea minutos después que como está configurado en el script ejecuta_llenar_datos.ksh, de esta manera con los datos obtenidos en el primer proceso se generará el archivo RDF.

Una vez que se tiene correctamente configurado el script, se procede a ejecutarlo desde la consola de linux por primera vez en background, de la siguiente manera:

4. Ubicarse en el path en donde se encuentra el script:
   cd wordpressmu/bookmarks/

5. Ejecutar el script en background:
   ksh nohup ejecuta_genera_rdf.ksh&

6. Validar de que el script se encuentre corriendo:
   ps -ef | grep genera | grep -v grep

   El resultado de ejecutar este último comando debe ser similar al siguiente:
   root 12075 10821 0 23:18 pts/1 00:00:00 kshejecuta_genera_rdf.ksh

   En donde **12075** es el Id del proceso en Linux que se está ejecutando correctamente en forma silenciosa.

Si por algún motivo se necesita cambiar algún parámetro de configuración del script, lo que se debe hacer es matar el proceso actual de Linux, cambiar la configuración y nuevamente ejecutar los tres pasos anteriores.

Para matar el proceso actual ejecutamos lo siguiente:

3. Obtenemos el Id del proceso actual:
   ps -ef | grep genera | grep -v grep

Debe arrojarnos un resultado similar al siguiente:
root 12075 10821 0 23:18 pts/1 00:00:00 kshejecuta_genera_rdf.ksh

4. Matamos el proceso:
   kill -9 12075

A continuación se muestra la pantalla de los archivos del sistema EMEB, con los archivos se generaran por cada mes y serán almacenados en la carpeta log.

Figura 75. Carpeta de archivos de logs generados por el sistema EMEB

La información que se registra en los archivos logs, cuando el sistema ejecuta los procesos automáticos como: el número total de bookmarks encontrados, bookmarks generados a RDF, así como también la fecha, hora de inicio y fecha en la que finalizó. Ejemplo:

El archivo LogLlenarDatos_201008.log

```
\n\nEjecutando archivo
Fecha Inicio: 2010/08/13 23:20:00
  No existen bookmarks nuevos en los blogs.Fecha Fin      :
2010/08/13 23:20:04
\n\nEjecutando archivo
Fecha Inicio: 2010/08/13 23:40:01
  Total bookmarks nuevos encontrados: 9Fecha Fin    : 2010/08/13
23:40:36
\n\nEjecutando archivo
Fecha Inicio: 2010/08/14 01:17:01
  Total bookmarks nuevos encontrados: 9Fecha Fin    : 2010/08/14
01:17:35
```

El archivo LogGeneraRdf_201008.log

```
\n\nEjecutando archivo
Fecha Inicio: 2010/08/13 23:22:01
  Total de bookmarks tomados para generar archivo RDF:  3
Fecha Fin    : 2010/08/13 23:22:04
\n\nEjecutando archivo
Fecha Inicio: 2010/08/13 23:45:01
  Total de bookmarks tomados para generar archivo RDF:  9
Fecha Fin    : 2010/08/13 23:45:03
```

```
\n\nEjecutando archivo
Fecha Inicio: 2010/08/14 01:18:02
  Total de bookmarks tomados para generar archivo RDF:  9
Fecha Fin    : 2010/08/14 01:18:03
```

3. CONSULTA DE LOGS

3.1 Tipos de logs utilizados

El uso de logs hace que el sistema de logs de un ordenador sea fundamental, ya que puede ocurrir cualquier anomalía que presente el sistema EMEB, o la misma base de datos del wordpress. Por tal razón en este sistema, se debe guardar o dejar un rastro, un comentario sobre lo ocurrido, en un fichero de registro, que permita poder solucionar el problema y así mantener informado al administrador de los procesos que se hayan ejecutado.

Se emplean 4 tipos de logs como:

ERROR: Trata a nivel de error, ejemplo: error en la inserción de algún bookmark.

INFO: Trata de la información más relevante (poco)

BELONG: Trata de errores con un nivel más alto de detalle del log ocurrido.

WARNING: Errores del tipo advertencia como: Exceder el numero de caracteres o por no poder sacar tags, mediante la evaluación mediante le web service Alchemy API.

En la Figura 76 es la interfaz de la consulta de logs, mediante la cual se escoge la fecha y el tipo de log para filtrar.

CONSULTA DE LOGS

Fecha Desde: Fecha Hasta: Tipo de Log: INFO

Consultar Logs

Detalle de logs desde el 23/08/2010 hasta el 23/08/2010, tipo: Todos.

Fecha	Tipo de Log	Descripcion	Pagina	Blog	Post
23/08/2010 09:19:29	INFO	INICIO PROCESO DE GENERACI	bk-llenar-datos.php		
23/08/2010 09:19:29	DEBUG	Inicia Leyendo Blog	bk-llenar-datos.php	Blog de Arquitectura	
23/08/2010 09:19:29	DEBUG	Obteniendo informaciÃ³n del Post: PresentaciÃ³n de Slide Share	bk-llenar-datos.php	Blog de Arquitectura	PresentaciÃ³n de Slide Share
23/08/2010 09:19:29	DEBUG	Bookmark insertado	bk-llenar-datos.php	Blog de Arquitectura	PresentaciÃ³n de Slide Share
23/08/2010 09:19:29	DEBUG	Obteniendo informaciÃ³n del Post: Agua mineral	bk-llenar-datos.php	Blog de Arquitectura	Agua mineral
23/08/2010 09:19:29	WARNING	DescripciÃ³n muy corta para generaciÃ³n de logs automÃ¡ticos	bk-llenar-datos.php	Blog de Arquitectura	Agua mineral
23/08/2010 09:19:29	DEBUG	Bookmark insertado	bk-llenar-datos.php	Blog de Arquitectura	Agua mineral
23/08/2010 09:19:29	DEBUG	Fin Leer Blog	bk-llenar-	Blog de	

Figura 76. Interfaz de la consulta de logs generados

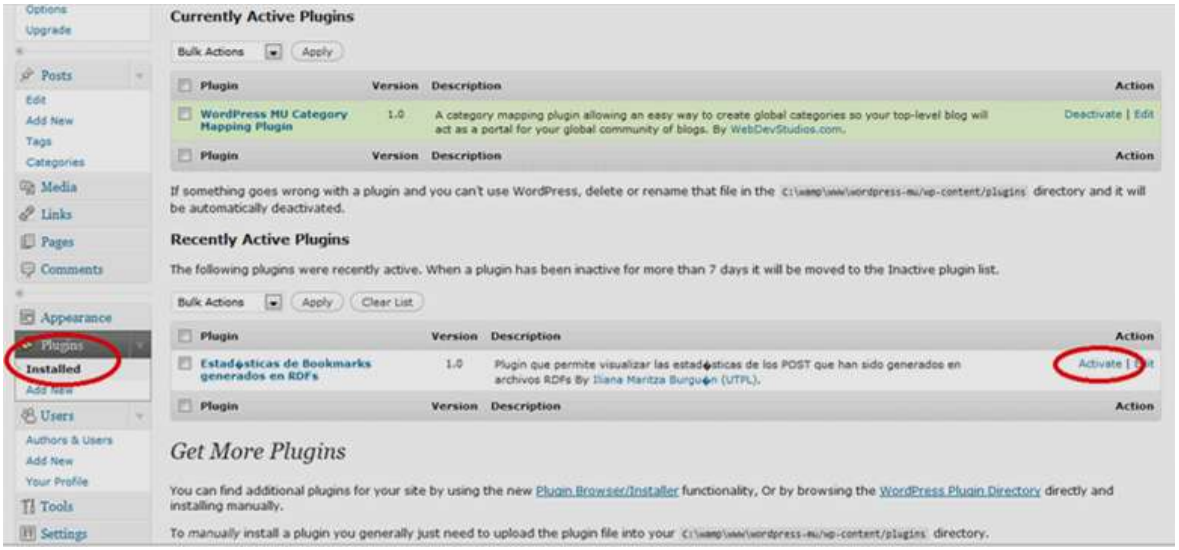
ANEXO 16: INSTALACIÓN Y USO DEL PLUGÍN BOOKMARKS EN RDFS

Para llevar el resumen de los blogs y entradas que han sido generadas a RDF, se emplea el plugín

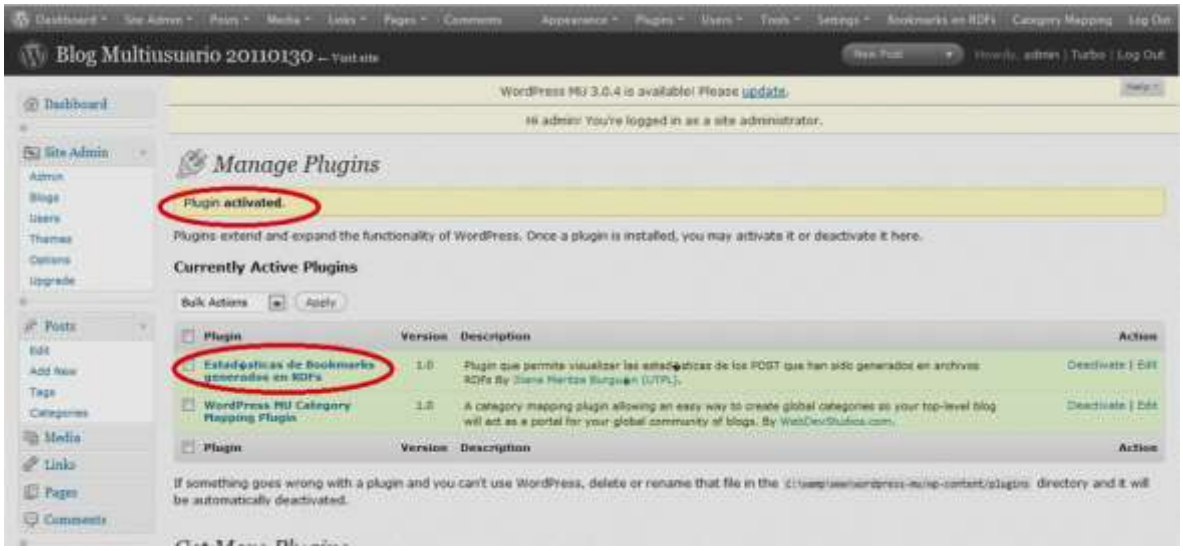
Bookmarks en RDFs.

El proceso de instalación es el siguiente.

1. Mediante el panel de administración del wordpress multiusuario, cargar el archivo mediante la pestaña "Plugins".
2. Pulsar en "Add New", escoger la ruta del archivo del plugin Bookmarks en RDFs.
3. Luego pulsar en "Installed", lugar donde se encuentran los plugins.
4. Revizar los "Recently Active Plugins", ubicar el nombre de "Estadísticas de Bookmarks generados en RDFs", pulsar "Activar".

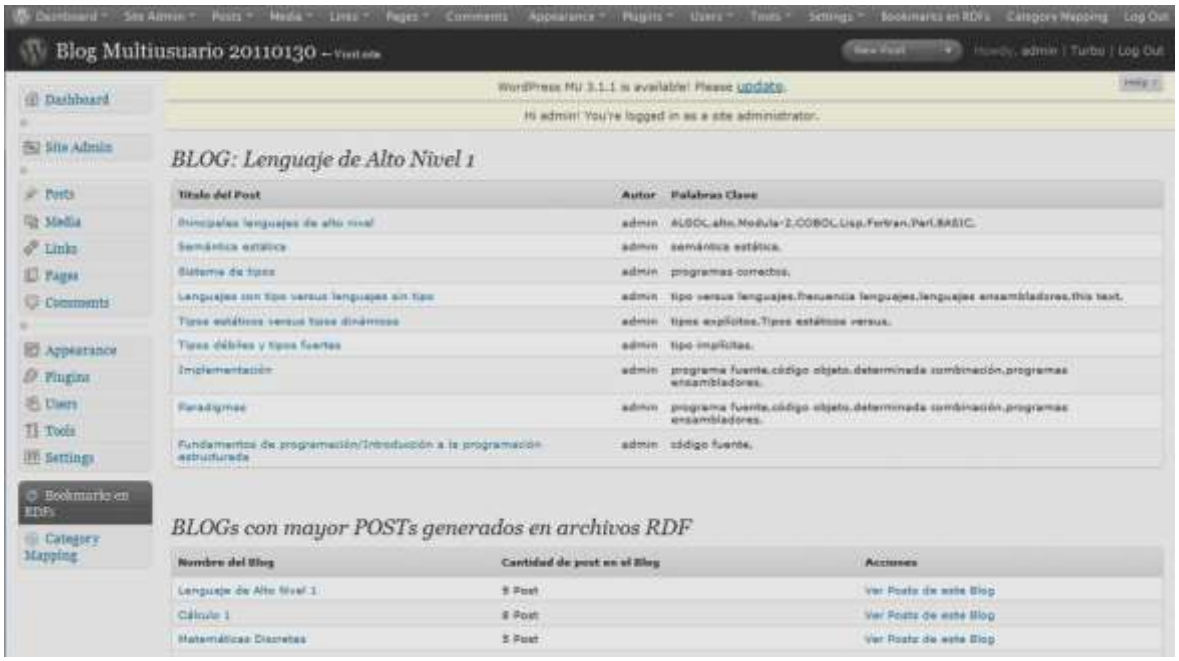


5. Se mostrara un pestaña como:



6. Luego que los procesos automáticos como "Llenar datos" y "Generar archivos RDF" hayan sido terminados, se registra en el plugín, el mismo que mantendrá informado al administrador de los 15 primeros blogs con mayor post generados en archivos RDF.

7. Así como también se puede ver los post, el blog al que pertenece, el titulo de post, usuario y sus palabras clave (tags).



ANEXO 17. TEST DE ESTABILIDAD DEL SISTEMA EMEB

Sr. Usuario el presente test tiene como objetivo evaluar la estabilidad del sistema EMEB, por el cual pido su contribución sírvase ingresar una entrada con los siguientes parámetros:

Ingrese al panel de administración de wordpress de pruebas en el sitio web http://blogsprueba.utpl.edu.ec/wp-admin/

Ingrese con la siguiente cuenta de usuario:

User: *pruebas*
Password: *pruebas*

Caso contrario cree un nuevo usuario

1. **Escoja cualquiera de los siguientes ítems.**

a) Ingrese una entrada de texto con etiquetas
   i. Escriba el titulo que ingreso
   ……

b) Ingrese una entrada de texto sin etiquetas
   i. Escriba el titulo que ingreso
   ……

c) Ingrese una entrada con un archivo y etiquetas
   i. Escriba el titulo que ingreso
   ……

d) Ingrese una entrada con un archivo y sin etiquetas
   i. Escriba el titulo que ingreso
   ……

e) Ingrese una entrada sin contenido
   i. Escriba el titulo que ingreso
   ……

f) Ingrese una entrada sin categorías
   i. Escriba el titulo que ingreso
   ……

g) Ingrese una entrada con código embebido de redes sociales como (youtube, slideshare, flickr)
   i. Escriba el titulo que ingreso
   ……

2. **La entrada que ud. ingreso se publico normalmente**
   Si …… No…..

ANEXO 18: EVALUACIÓN DE TAGS LIBRE DE CONTEXTO

- ✓ Tags ingresados por usuarios
- ✓ Tags extraídos por Alchemy API
- ✓ Umbral de operaciones de la aplicación de la Distancia de Levenshtein
- ✓ Valores del experimento 1
- ✓ Valores del experimento 2

El Wordpress multiusuario como su nombre lo dice es utilizado por muchos usuarios, los mismos que tratan diversos temas. Para la aplicación de esta prueba se divide en dos grupos.

1. Tags ingresados por usuarios

seminario contabiildad, congreso, internacional, bursátil, invesigación, proyecto, bolsa, película, wall street, Banca y Finanzas, Matriculas, UTPL, Avisos, Calendario, Cariamanga, Exámenes Presenciales, UTPL, Cariamanga, ECC, Semana Informática, UTPL, Cariamanga, Emprender, Estudiantes, Gestión Productiva, MIPRO, Relaciones Interinstitucionales, Servicios, UTPL, Banco Mundial, Cambio Climático, Cariamanga, GDLN, Servicios, UTPL, Cariamanga, Cultura, Cariamanga, Docencia, Estudiantes, Misiones Universitarias, Cariamanga, CITTES, ECTS, Estudiantes, Gestión Productiva, UTPL, Cariamanga, Donación, Estudiantes, Gestión Productiva, Informática, Labor Social, Relaciones Interinstitucionales, UTPL, Cariamanga, Conferencias, Docencia, Estudiantes, Servicios, SRI, UTPL, Cariamanga, Donación, Relaciones Interinstitucionales, Servicios, UTPL, Cariamanga, Estudiantes, Juventud Idente, Labor Social, UTPL, Cariamanga, Emprender, Estudiantes, UTPL, Cariamanga, Conferencias, Informática, UTPL, El ADN, biodiversidad, biodiversidad Ecuador, genética, AULA VIRTUAL, CENTRO PROVINCIAL RIOBAMBA, Noticias, Computación Básica, Computación Básica, Computación Básica, Computación Básica, Computación Básica, Computación Básica, Computación Básica, Computación Básica, Computación Básica, Computación Básica, UTPL, Web Semántica, Dries, EVOC, RDF, SIOC, UTPL, EMBED, Validar XHTML, W3C, YOUTUBE, DERI, Drupal, Obama, recovery.org, SIOC, DERI, RDF, SIOC, WEB, Web Semántica, UTPL, Web Semántica, distancia, educación, iped, utpl, Convenios, convenioutpl, Entrevista Redes sociales UTPL Loja, CEDIA, CEPRA, ENSEÑANZA, PUCE SD, PUCE SI, RED, Reunión RVEA, UTPL, Videoconferencia, CEDIA, CEPRA, ENSEÑANZA, PUCE SD, PUCE SI, RED, Reunión RVEA, UTPL, Videoconferencia, Directiva videoconferencias utpl 2010, Política exterior Videoconferencia UTPL, ética, Periodismo, UTPL, Videoconferencia, E-Learning, Videoconferencias Internet, Añadir etiqueta nueva, oferta, Día Bibliotecario, facebook, marketing 2.0, Word 2010 - guía, !Phone 3GS, animación, AutoCad, fotografía, fotomontaje, google, Libros computación, Modelado Humano 3D, Presto 10.2, Windows 7, windows server 2008, Histología, Sobotta, segmentación, paginación, memoria, Añadir etiqueta nueva, Paralelo "A", Gestion Memoria, linux, sistemas operativos, Añadir etiqueta nueva, Paralelo "A", Planificacion procesos, maquinas virtuales, mandriva, snow leopard, windows 7, gestion memoria, maquinas virtuales, peor y mejor ajuste, primer, Conexión Remota, Recuperación archivos, Sacar respaldo, Transferencia remota archivos, Directorio, Fichero, Termina, Pipewalker, acceso remoto, herramientas, respaldo archivos, restaurar archivos, procesos, Utpl, Conferencia, opinion, hilos, Seccion Critica, sincronizacion, interbloqueso procesos, universidad tecnica particular loja, Utpl, Dr. Orihuela, Utpl, Planificacion procesos, Utpl, linux solaris XP, planificacion, algoritmos planificacion, universidad tecnica particular loja, Utpl, Planificacion CPU, Utpl, linux, planificacion sistemas operativos, planificacion SO, sistemas operativos, SO, solaris, windows XP, Planificacion CPU, Capitulo 5, planificacion CPU, resumen, CPU, corto plazo, largo plazo, mediado plazo, planificadores, proceso, procesos, sistemas operativos, Utpl, procesos sistema operativol, universidad tecnica particular loja, Utpl, evaluaciones, postgrados, campaña utpl 2.0, UTPL, Carlos Vera, comunicación, día, digital, Ecuador, internet, Loja, periodismo, social, UTPL, Metodología, Programación II,

Total de tags: 260
Palabras eliminadas: Por, del, de, la, en, los,

2. Tags extraídos por Alchemy API

Contabilidad, Bancaria, Universidad Técnica Particular, Matriz Campus UTPL, asistencia ingresa, libre configuración, Escuelas Administrativas, opiniones, Wall Street, magnate manipulador, Michael Douglas, verdadera realidad, Gordon Gekko, MATRICULA, FINANCIERO ECUATORIANO FRENTE, ECONOMÍA ACTUAL, cordial invitación, Aula Magna, Verónica Cuenca, ASO, Cetro Universitario, créditos ects, Semana Informática UTPL, Informática UTPL Cariamanga, Instituto Tecnológico, Eloy Alfaro, Mariano Samaniego, Redes Inalámbricas, día miércoles, Centro Universitario, día viernes, plataforma base, créditos utpl-ects, libre configuración, Además Conferencias, Mantenimiento Preventivo, Web Semántica, Semana Informática UTPL, Informática UTPL Cariamanga, concurso emprender, Instituto Tecnológico, Ricardo Donoso, Mauricio Eguiguren, extensión universitariaRicardo Donoso coordinador, Universidad Técnica Particular, Eloy Alfaro, Mariano Samaniego, cantón calvas, Productividad MIPRO, Centro Universitario, día miércoles, Redes Inalámbricas, créditos utpl-ects, Angel Soto, acreditación utpl-ects, día viernes, Además Conferencias, Web Semántica, Jefe Político, libre configuración, micro talleres, banco mundial, Juan Manuel García, vecino cantón, Fausto López, día jueves, Cambio Climático, abogada ximena torres, Magister Ricardo Donoso, artístico sei corde, solemne acto, XXVIII aniversario, siglo xx, Extensión Universitaria, Gestión Cultural, Reflexión Espiritual correspondiente, José María López, Extensión Universitaria, día jueves, verdadero sentido, verdadera familia, 8vos ciclos, Extensión Universitaria, María Jiménez, Gestión Productiva, perfectas condiciones, Escuela Juan, barrio lanzaca, valioso apoyo, Extensión Universitaria, Instituto Pedagógico Ciudad, importante material, importante aporte, Nancy Castillo, Extensión Cariamanga, UTPL Extensión Cariamanga, Infantería Capitán Díaz, Escuela San José, Cantón Calvas, sencillo ágape, destrezas motrices, Juveniles Identes, día venideros, Universidad Técnica Particular, iniciativa empresarial, UTPL Cariamanga, formación académica, población calvense, Pablo Torres, Juan Pablo Pardo, Juan Paúl Jiménez, 8vo ciclo, Marco Vivanco, cincuenta estudiantes, día viernes, inglés deoxyribonucleic acid, virus adn, ARN polimerasaTAC-GAT-CTA-GCG-, grupo fosfato, largo trenvagones, ARNm resultante, largas cadenas, organismos eucariotas, organismos procariotas, vista geográfico ecuador, Río Napo, Cordillera Occidental, Cordillera Oriental, zonas marítimas pesqueras, medio ambiente marino, Biodiversidad Tiputini, prolífica población, ríos anchos, proverbial diversidad, corrientes frías, exuberante vegetación, inmediato contorno, extensas planicies, prolífica flora, Cuenca Amazónica, Puerto Bolívar, maravillosos mundo, selvasEl ecosistema, formando suelos, temperatura promedio, numerosos valles, peculiar formación, múltiples formas, extraordinarias variaciones, excelentes lugares, principal atracción, proteínas orden, Gregor Mendel, siglo xix, AULA VIRTUAL, Nueva Loja, siguientes centros, Quito-San Rafael, centro provincial riobamba, históricas primicias, Calle Juan Chiriboga, Universidad Técnica Particular, majestuosa ciudad, centro machala, Circunvalación Norte esq, utpl centro machala, Madero Vargas, único campo, referencias relativas, celda a7, celda a6, Función BUSCARV, Función BUSCARH, FUNCIÓN PROMEDIO, FUNCIÓN SI, herramienta wordart, Tablas bordes, ABRIR DOCUMENTOS, palabras word, EL ECUADOR, accesibilidad web wai, Marx Ortiz, web semántica, portal universitario, datos vía rdfa, Drupal 6.x, CMS Drupal, Web semantica, Dries Buytaert, módulo rdf, formato rdf, módulo evoc, XHTML Strict, codigo youtube, ultimo doc type, Web Semántica, Web Semántica, EE.UU Barack Obama, CMS Drupal, Web Semántica, Enterprise Research Institute, Universidad Técnica Particular, famosas redes, formato rdf, web semántica, usuario respuestas específicas, ranking y posicionamiento, Oswaldo Barrera, Donato Vallín González, Universidad Técnica Particular, Luis Jaramillo Pacheco, GRANA Lanzamineto, Acosta Santiago, Willie Moreno, Miskulin

Mauro, América Latina, Junta Directiva, Unidad Educativa, Unidad Educativa Rubén, Radio Municipal, Unidad Videoconferencias-UTPL, José Luis Granda, coordinador unidad videoconferencias-utpl, social educativa eqaula, día miércoles, Jorge Guamán, GDLN América, red cedia, red virtual, santo domingo, plataforma moodle glesone, Pontificia Universidad Católica, Universidad Técnica Particular, sede santo domingo, universidades participantes, Consorcio Ecuatoriano, Milton Andrade, Franklin Sánchez, sede ibarra, proyecto rvea, VIRTUAL DE, Reunión Proyecto RVEA, INTEGRAL TECNOLOGÍAS, RED VIRTUAL DE, Universidad Técnica Particular, NACIONAL VIRTUAL DE, presente proyecto, Espacios Físicos, Nueva Directiva, Jorge Guamán, Fabricio Paredes, Arturo Valenzuela, Hemisferio Occidental, Universidad Técnica, Estados Unidos, presidente clinton valenzuela, Exterior Edmund Walsh, José Barbosa Corbacho, Universidad Técnica Particular, Secretario Adjunto, principales asuntos, Subsecretario Adjunto, rector canciller, mejores tácticas, Casa Blanca, Universidad Técnica Particular, PERIODISTICA TIEMPOS, Nuevo Periodismo Iberoamericano, Xavier Darío Restrepo, Autoregulación Periodística Iberoamericana, Andina Simón Bolívar, María Alfaro Moreno, Jornadas Periodísticas, Open Society Foundation, Directora Ejecutiva, Directora Legal, Kela León, Cynthia Cárdenas, presente año, EXPERTOS NACIONALES, Llucía Oliva, temas claves, Artículo XXI, open source, power point, MEETING OPEN SOURCE, plataformas open source, comunicación open source, server red5, servidor red5, modelo e-learning, tareas descritas, Remote Desktop, Giorgio Natili, Shared Desktop, pequeño ajuste, documentos ppt, Casa Abierta UTPL, Jorge Luis Borges, bibliotecario ecuatoriano, Eugenio Espejo, biblioteca pública, ilustre personaje eugenio, Luis Gallo Porras, Febrero Día, Proverbio Indú, delicado elogio, historia ecuatoriana, Emily Dickinson, Ernest Steinbeck, redes sociales, FACEBOOK EXPRIME, Novedades periodo agosto, NUEVO MARKETING EN, FENÓMENO MASAS, Juan Manuel Maqueira, REDES SOCIALES, sume adeptos, SIGUIENTES LIBROS, siglo xx, darán alternativas, GUÍA PRÁCTICA, Sebastián Bruque, Estados Unidos, instrucciones paso, Photoshop CS4, Windows Server, iPhone 3GS, herramientas 3D, Photoshop CS4 Extended, aplicación google earth, Web Google Chrome, potentes herramientas 3D, Google Talk, Google Reader, mejores fotomontajes, Dé rienda, Hyper-V Server, Gran Muralla, manual impreso, Anaya Multimedia, tecnología hyper-v, rotundo éxito, elegante interfaz, nivel anfitrión, BIBLOTECA, David Pogue, excelente aprendizaje, potente programa, múltiples ubicaciones, única ventana, Windows Vista, excelentes resultados, elegante ordenador, llamada biblioteca, usuario multitud, excelentes explicaciones, Joe McNally, prestigiosas revistas, mejores fotógrafos, humano 3D, cuadro multicolor, segmentos ro, llamadas marcos, llamadas páginas, memoria dinámica, área libre l, partición p, Peor Ajuste, Mejor Ajuste, P_TAMAÑO= L_TAMAÑO, tamaño p_tamaÑo, P_BASE= L_BASE, bloque l, lista, memoria, memoria ram, principales recursos, altos requerimientos, alta capacidad, Software Libre, Richard Stallman, Linus Torvalds, Linus Torvalds Linux, Kernel Linux, Linus Benedict Torvalds, actual kernel linux, operativo libre gnu, interesante sistena operativo, Unix Operating System, GNU Compiler Collection, Bourne Again Shell, Source Development Labs, Licencia Pública General, ideología comunista, Josselyn Arias, sistema gnu, compilador gcc, tanta frecuencia, siguientes operaciones, computadoras multiprocesador, maquina virtual, maquinas virtuales, Red Hat, Red Hat Enterprise, Sistema Operativo CentOS, comando rm, CentOS usa yum, ENTerprise Operating System, Hat Enterprise Linux, comando numero, comando touch, comando history, touch myfile1 myfile2, –p file1 file2, comando cp, chmod g+x carta, certeza qu, programa simula, único ordenador, opción argumento, chmod u=rwx, Memoria RAM, touch myfile2, cal mes, código fuente, -r enlista, carpetas llamadas, cp, nombre ruta, go=rw carta, Snow Leopard, MÁQUINAS VIRTUALES, Mejor ajuste, Peor ajuste, RECUPERACIÓN ARCHIVOS, REMOTA DE, sistema operativo centos, usuario root, máquina virtual vmware, vía comandos, presento programas, única hebra, os procesos, procesos creadores, procesos padre, procesos hijos, procesos.-un proceso, determinado suceso, bloque control, largo plazo, corto plazo, Pregunta curiosidad, PROCESOS, semáforo contador, procesos estén, semáforo s, múltiples procesos, correspondiente sección, procesos

escritores, semáforos contadores, Roberto Valladolid, correspondientes secciones, cola fifo, variable entera, tipo monitor, búferes vacios, esquema lifo, principal desventaja, soporte hardware, maquina smp, amplia clase, INTERBLOQUEOS, recurso cpu, ASIGNACION RECURSOS, sistema operativo información, conjunto {Po, Jorge Luis Orihuela, libre expresión, ideas, corto plazo, planificación prioridad, SO, hebra especial llamada, clase sistema, colas multinivel, Linux asigna, tareas caducas, matriz caduca, Windows XP, prioridad mas, windows xp, tiempo, Sistema Operativo Solaris, planificación gracias, windows xp,PLANIFICACION DE, PLANIFICACION SO, PLANIFICACION SOLARIS, Marilin Zárate, denomina despachador, windows xp, sistemas operativos, prioridad, sistemas operativos solaris, kernel llamada despachador, procesos interactivos frente, Karina Jimenes, CLASE VARIABLE, planificación fcfs, DECPU, PLANIFICACION LA, Planificacion CPU, procesos interactivos frente, planificación sjf, colas multinivel, planificación sfj, procesos linux, algoritmo sjf, algoritmo fcfs, turnos degenera, SILBERSCHATZ GALVIN GAGNE, corta duración, múltiples procesadores, Santiago Aguilera, Windows XP, colas multinivel, Planificación SJF, algoritmo sjf, Planificación FCFS, planificación mas, corta duración, menor tiempo, trabajo, Windows XP, Windows XP, cpu,ejecución, sistemas multiprocesador, Planificación FCFS, Planificación SJF, sistemas operativos solaris, Cesar Capa González, corto plazo, prioridad, colas multinivel, pequeña unidad, trabajo, corto plazo, medio plazo, largo plazo, Windows XP, Roberto Valladolid, proceso hijo, proceso padre, proceso hijo carga, remote method invocation, largo plazo, memoria compartida, memoria múltiples programas, sockets udp, sockets tcp, esquema rpc, mecanismo rpc, socket multifunción, clase socket, adecuada llamada, única hebra, IPC interprocess, siguientes razones, corto plazo, soporte hardware, arquitectura cliente-servidor, clase datagramsocket, dirección ip, clase multicastsocket, espera proceso, memoria múltiples programas, procesos, sistema exit, ultima instrucción, llamada rpc, determinado puerto, socket cliente, primeras, revisar mayor, Información Postgrados, serán, More Info Url, videos, actividades, Estudiante , etiqueta utpl, palabra utpl, UNIVERSIDAD CUENTA CONTIGO, categoría utpl, Usa TAGS, Ranking Webometrics, siguientes semanas, Dirección General, sitio utpl, ¿Eres usuario, universidad www.utpl.edu.ec, Campaña Docentes, extensión pdf, Campaña Estudiantes, Comunicación social, Ecuador, géneros periodísticos, Loja, UTPL, periodismo digital, CONCURSO PERIODISMO DIGITAL, Catalina Mier, caso catalina mier, Punín Docente investigadora, Jurado Calificador, Cobertura Digital, Espinoza Blogger, Friend Feed, Programación, Metodología, Programación II, Universidad Técnica Particular Loja, Ciencias, Patricio Abad Espinoza, estado,
Total de tags: 625
Palabras eliminadas: en, los, el, de, La, la, las, que, mas, tus, una, —el, *La

3. Aplicando el Algoritmo de la Distancia de Levenshtein, se adquiere un umbral de promedio de: 5

Tags ingresados por usuarios	Tags extraídas por Alchemy API	Distancias
Matriculas	MATRICULA	1
Semana Informática	Semana Informática UTPL	4
Emprender	concurso emprender	8
Cariamanga	UTPL Cariamanga	4
biodiversidad	Biodiversidad Tiputini	8

EVOC	modulo evoc	6
YOUTUBE	codigo youtube	6
Drupal	CMS Drupal	3
CEDIA	red cedia	3
Reunión RVEA	Reunion Proyecto RVEA	8
facebook	FACEBOOK EXPRIME	7
segmentación	segmentos ro	5
peor mejor ajuste	mejor ajuste	5
Planificacion CPU	PLANIFICACION	3
procesos	proceso hijo	5
UTPL	sitio utpl	5
periodismo	periodismo digital	7

4. Para la obtención de los falsos positivos y falsos negativos se evalúa contra una extracción de tags de Misiones universitarias.

Entrada evaluada	**Tags extraídos mediante alchemy api**
http://blogs.utpl.edu.ec/misionescariamanga/	Misión Idente Ecuador, Universidad Técnica Particular de Loja, Instituto Id, Cristo Redentor, actividad misionera, Humanismo Cristiano, Misiones Universitarias, Misioneros Identes,

La siguiente tabla resume los resultados obtenidos aplicando el algoritmo de Distancia de Levenshtein:

Verdaderos positivos	Verdaderos negativos	Falsos positivos	Falsos negativos
43	220	2	6

5. Para comparar sus resultados se evalúa contra la misma entrada del literal 4.Cuyos resultados sin aplicar algoritmos son:

Verdaderos positivos	Verdaderos negativos	Falsos positivos	Falsos negativos
23	236	2	6

BIBLIOGRAFÍA

[1] Marzal M., García Q., Butera M., Los blogs en el nuevo modelo educativo universitario: posibilidades e iniciativas., Disponible en: http://www.ub.es/bid/19marza2.htm

[2] Lara T., La utilidad de un blog académico., Disponible en: http://tiscar.com/2006/09/14/la-utilidad-de-un-blog-academico/

[3] Sonia SÁNCHEZ-CUADRADO, Juan LLORÉNS, Jorge MORATO y José A. HURTADO., Extracción Automática de Relaciones Semánticas, Departamento de Informática, Universidad Carlos III Leganés (Madrid), España, Disponible en: http://www.iiisci.org/journal/risci/Abstract.asp?var=&id=P338146

[4] Nov. 2009. Martínez Felipe. Memoria del IV Encuentro de Catalogación y Metadatos. Disponible en: http://132.248.242.3/~publica/archivos/libros/iv_encuentro_catalogacion.pdf

[5] Shangai, China, Intenational Conference on Dublin Core and metadata Applications 2004., "Metadata for Interoperability in the Global Corporate Environment". Disponible en: http://dc2004.library.sh.cn/english/prog/pro-co.htm

[6] Manola Frank., Miller Eric., W3C Recommendation. RDF Primer. Disponible en: http://www.w3.org/TR/rdf-syntax/#rdfmodel.

[7] Lamarca M., "RDF" Disponible en: http://www.hipertexto.info/documentos/rdf.htm

[8] Stanford Univerity, InfoLab., Simplified Syntax for RDF. Disponible en: http://infolab.stanford.edu/~melnik/rdf/syntax.html

[9] Uso del DublinCore (DCMI). ISO 15836-2003. Disponible en: http://www.sedic.es/autoformacion/metadatos/tema7.htm

[10] Wikipedia., Dublín Core., Disponible en: http://es.wikipedia.org/wiki/Dublin_Core

[11] WordPress Codex DataBaseDescription., Disponible en: http://codex.wordpress.org

[12] Precision and recall. Disponible en: http://en.wikipedia.org/wiki/Precision_(information_retrieval)#Definition_.28classification_context.29

[13] Snowball. Disponible en: http://snowball.tartarus.org/algorithms/spanish/stop.txt

[14] Distancia de Levenshtein., Disponible en: http://es.wikipedia.org/wiki/Distancia_de_Levenshtein

[15] Recuperación y organización de la información, Universidad Carlos III de Madrid. Disponible en: http://recuperacionorganizacioninformacion.50webs.org/

[16] Borjars U., Breslin J., Moller K., Using Semantics to Enhance the Blogging Experience., Digital Enterprise Research Institute, National University of Ireland. Disponible en: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.104.5396

[17] What would it mean to blog on the semantic web., David R. Karger , Dennis Quan Disponible en: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.97.3624

[18] SIOC-PROJECT.ORG. Disponible en http://sioc-project.org/

[19] MORFEO PROJECT., Técnicas de descubrimiento de recursos en función del contexto., Disponible en: http://forge.morfeo-project.org/wiki/index.php/D.4.1_T%C3%A9cnicas_de_descubrimiento_de_recursos_en_funci%C3%B3n_del_contexto

[20] Anieto2k., Disponible en: http://www.anieto2k.com/2007/08/27/taxonomia-en-wordpress-23/

[21] Esquema taxonómico de WordPress., Disponible en: http://www.yukei.net/2007/11/esquema-taxonomico-de-wordpress/

[22] Foca Online, Servicio web disponible para la extracción de metadatos en recursos. Disponible en: http://www.informatica64.com/foca/

[23] Soft Experience, Metadata Miner Catalogue PRo software, Disponible en: http://www.metadataminer.com/

[24] Soft Experience, Metadata Miner Catalogue Pro software, Disponible en: http://peccatte.karefil.com/Software/Software.html

[25] Calais. Disponible en: http://www.opencalais.com/about

[26] Wordpress.org., Plugin Directory., Disponible en: http://wordpress.org/extend/plugins/tagaroo/

[27] Sitio oficial de Alchemy API http://www.alchemyapi.com/api/demo.html

[28] Manuel Fernando Caro, Caracterización de los objetos digitales de aprendizaje elaborados como producto de las actividades de docencia y extensión en la facultad de educación y ciencias humanas de la Universidad de Córdoba, Disponible: http://edupmedia.org/index/descargas/ARTICULO-CACUMEN-VOL2.pdf

[29] Adriana J. Berlanga1, Clara López1, Erla Morales2, Francisco J. García1

1Departamento de Informática y Automática, Universidad de Salamanca, España "Consideraciones para Reforzar el Valor de los Metadatos en los Objetos de Aprendizaje". Disponible en: www.uoc.edu/symposia/spdece05/pdf/ID03.pdf

[30] Red Iris., Disponible en: http://www.rediris.es/rediris/index.html.es

[31] A tool for creating Dublin Core metadata., Disponible en: http://www.ukoln.ac.uk/metadata/software-tools/

[32] 68th IFLA Council and General Conference August 18-24, 2002. Disponible en: http://archive.ifla.org/IV/ifla68/papers/067-152s.pdf.

[33] University of Cambridge., Disponible en: http://www.cam.ac.uk/cs/web-search/metatags.html

[34] Guidelines for Use of Dublin Core in University of Chicago Library Projects Disponible en: http://memory.loc.gov/ammem/award99/icuhtml/dcguide.html#3

[35] Theoretical Considerations of Lifecycle Modeling: An Analysis of the Dryad Repository Demonstrating Automatic Metadata Propagation, Inheritance, and Value System Adoption., Disponible en: http://www.informaworld.com/smpp/ftinterface~content=a910229369~fulltext=713240928~frm=content