**SERVICIOS WEB PARA EXTRACCIÓN DE ENTIDADES DESDE CONTENIDO HTML. PILOTO EN SITIOS CON RECURSOS ABIERTOS OCW**

Jhonny Zaruma, 2014

**Universidad Técnica Particular de Loja - Escuela de Ciencias de la Computación**

**INTRODUCCIÓN**

La web conforma un mudo de datos, información y conocimiento casi siempre nos encontramos con la dificultad de encontrar la información que realmente necesitamos. Existen muchos buscadores y algoritmos de búsqueda, pero falta mucho por recorrer para poder llegar a automatizar búsqueda y recuperación de información mediante búsquedas inteligentes.

Por lo tanto la asignación de metadatos y etiquetas es de gran importancia si se quiere tener una búsqueda inteligente en la web semántica. Con esto se encuentra la necesidad de una herramienta de extracción de meta información de calidad.

El presente trabajo se enfoca en el procesamiento de los recursos educativos abiertos OCW (Open Course Ware por sus siglas en inglés) que son una de las iniciativas educativas muy importante, ya que son libre acceso a una gran diversidad de recursos y materiales de cursos universitarios de forma gratuita e ilimitada.

**OBJETIVOS**

**GENERAL**

* Servicios Web que extraigan datos a partir del contenido HTML, el piloto se ejecutará sobre contenidos OCW.

**ESPECÍFICOS**

* Creación de WS para Tokenización
* Creación de WS para Extracción de entidades
* Creación de WS para Desambiguación y Limpieza
* Creación de WS para Enlace con LOD-Cloud

**CONCEPTOS GENERALES**

**PYTHON**

Python es un lenguaje de programación dinámico y potente que  es utilizado en varios dominios de aplicación . A Python se lo compara con Tcl, Perl, Ruby, Scheme o Java. Sus características mas relevantes son (Python):

* Su sintaxis es muy clara y legible
* Fuerte Capacidad de introspección
* Orientado a objetos
* Expresión natural del código procedimental
* Completamente modular, Soporte para paquetes jerárquicos
* Manejo de errores basado en excepciones
* Los tipos de datos son dinámicos de muy alto nivel
* Bibliotecas estándar y módulos de terceros para prácticamente todas las tareas
* Puede ser integrada en aplicaciones mediante interfaz de script

Python es un lenguaje de programación que te permite trabajar con mayor rapidez e integrar sus sistemas con mayor eficacia. También se puede ejecutar en Windows, Linux / Unix, Mac OS X, y ha sido adaptada a la de Java y máquinas virtuales. NET.

Python es libre de usar, incluso para los productos comerciales, por su aprobada por OSI licencia de código abierto. (Python)

**NLTK** (Nltk)**:** es una plataforma para python que trabajar con datos de lenguaje humano. Proporciona interfaces fáciles de usar, junto con un conjunto de bibliotecas de procesamiento de textos para la clasificación , tokenización , derivado, el etiquetado , el análisis y el razonamiento semántico.

NLTK se lo puede trabajar en Windows , Mac OS X y Linux. NLTK también es código abierto, impulsado por la comunidad.

El público objetivo de NLTK consiste en lingüistas y científicos de computación, y es accesible y desafiante en muchos niveles de las aptitudes computacionales. Y se basa en un lenguaje de programación orientado a objetos apoyándose prototipado rápido y programación literaria.. (Loper & Bird)

**SERVICIOS WEB**

Los servicios web(Web Service) son componentes software que permiten intercambiar información y datos entre aplicaciones, mediante el uso de tecnologías Web  basadas en estándares y protocolos. Los servicios web se diseñados para que se pueda acceder por otras aplicaciones. Los servicios web son un conjunto de herramientas que pueden ser usadas en distintas formas. (Alvarado Ruiz, Guamán Eras, & Sigcho Armijos)

Según el W3C (World Wide Web Consortium) los Servicios Web son aplicaciones de software identificadas por un URI (Uniform Resource Identiflier), cuyos interfaces y vínculos tienen la capacidad de estar bien definidos, descritos y descubiertos como objetos XML. WS soporta interacciones directas con otros agentes de software usando mensajes de intercambio basados en XML vía protocolos basados en Internet. Se puede perfeccionar esta definición pidiendo que la descripción se haga a través de un documento WSDL (Web Services Description Language) y el protocolo utilizado sea SOAP. (W3C)

Puesto que cada Servicio Web puede estar implementado en una tecnología heterogénea es necesario cumplir una serie de estándares para hacer posible la comunicación entre ellos. Los más utilizados son los siguientes (González):

* Web Services Protocol Stack: conjunto de servicios y protocolos de los servicios Web.
* XML: Es un lenguaje de marcas capaz de describir distintos tipos de datos. Es un estándar aceptado y utilizado como medio de descripción de datos.
* SOAP: Es un protocolo de comunicación entre procesos basado en el intercambio de mensajes en formato XML dentro de una red. A su vez SOAP está basado en XML y es completamente independiente de la plataforma y del lenguaje en el que estén implementados los procesos que se comunican.
* WSDL: Es un lenguaje basado en XML que permite describir servicios web (como su nombre indica). Un documento WSDL especifica, entre otras cosas, dónde se encuentra el servicio así como las operaciones que pone accesibles a otros servicios.
* UDDI: Es un directorio, basado en XML, en el que las distintas empresas dan de alta servicios web que ponen al servicio de otra empresas.
* WS-Security (Web Service Security): Protocolo de seguridad aceptado como estándar por OASIS (Organization for the Advancement of Structured Information Standards). Garantiza la autenticación de los “actores” y la confidencialidad de los mensajes enviados.

**Arquitectura Rest** (Alvarado Ruiz, Guamán Eras, & Sigcho Armijos)**:** Transferencia de estado representacional (REST) es un estilo de arquitectura de software para sistemas hipermedia distribuidos como la World Wide Web. El término fue introducido en la tesis doctoral en el año 2000 por Roy Fielding , y ha entrado en uso generalizado en la comunidad de redes.

REST se refiere estrictamente a una colección de principios de la arquitectura de red que describen cómo los recursos son definidos y tratados. El término se utiliza a menudo en un sentido más amplio para describir cualquier interfaz simple que transmite datos específicos del dominio a través de HTTP y sin capa de mensajería adicionales, tales como SOAP o una sesión de seguimiento a través de cookies HTTP.

**WEB SEMÁNTICA**

El creador del concepto, Tim Berners-Lee, define la web semántica de la siguiente manera: “*no es una web separada sino una extensión de la actual, donde la información está dotada de un significado bien definido, los ordenadores están mejor capacitados y las personas trabajan en colaboración*” (La Web Semántica y las Tecnologías del Lenguaje Humano)

****

Figura 1 Arquitectura de la web semántica (Fuente Tim Berners-Lee,  
http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html).

Un objetivo de la Web semántica es crear un sistema de agentes inteligentes que puedan hacer deducciones de una manera automatizada con la información que esta en la Web. Este objetivo más que una realidad es una utopía incluso a medio plazo. Por otro lado, los desarrollos que se han realizado gracias a este nuevo paradigma han dado lugar a nuevos servicios ajustados con éxito en la actual Web. Como por ejemplo se logrado construir diferentes estándares para poder representar y procesar la información de una manera mas sofisticada. Estos estándares que han permitido presentar los metadatos en un formato mas lógico y controlados (por ejemplo ontologías) para que sean procesados por programas informáticos. Estos formatos ya son utilizados de manera generalizada, como por ejemplo XML, RDF, SKOS-Core y OWL. (Vallez, Rovira, Codina, & Pedraza)

### METADATOS Y ANOTACIÓN SEMÁNTICA

Un elemento fundamental de la Web semántica son los metadatos, en otras palabras, información que nos describe el contenidos de los documentos a los que está ligado y nos representa de una manera explícita el significado de estos.

La anotación semántica echa con metadatos ofrece contenido semántico a los documentos para logra que las máquinas interpreten la información.

Las herramientas de anotación permiten convertir en metadatos el contenido semántico extraído de las páginas web. Existen herramientas de anotación dirigidas a los autores ayudan a incorporar los metadatos dentro o fuera de las propias páginas web siguiendo los estándares (xml, rdf...). Y Las aplicaciones del herramientas de anotación externa permiten asociar metainformación a páginas web, pero esta no se almacena dentro de la misma página sino que se guardada de forma externa en un repositorio.

Existen diferentes aproximaciones para realizar la anotación semántica, pero se pueden agrupar en tres grandes categorías. (Vallez, Rovira, Codina, & Pedraza)

* El primer modelo se basa en la anotación lingüística, el objetivo es etiquetar los textos a partir de los diferentes niveles de la lengua. Resulta de gran interés la identificación de los términos y saber cómo estos se relacionan entre sí porque esta información puede incidir en el valor de un término como palabra clave. Este sistema es muy costoso computacionalmente.
* La segunda aproximación se basa en las ontologías, son utilizadas como recurso principal para extraer las conexiones entre los términos y representar su significado.
* La tercera aproximación propone el uso de un lenguaje controlado, este es un modelo que está directamente vinculado con la asignación de metadatos y la anotación semántica.

**OCW**

OpenCourseWare (OCW) es un ejemplo de las iniciativas que desde el 2001 han emergido para promover el acceso libre y sin restricciones al conocimiento.

(UTPL, 2014)

**OCW site** (UPM)**:** Es un espacio web que contiene materiales docentes creados por profesores para la formación superior. Las características que distinguen al proyecto OpenCourseWare de iniciativas similares son las siguientes:

* Los recursos didácticos publicados en un OCW site se organizan en unidades de “asignaturas” o “cursos”. Con ello se quiere indicar:
  + Los accesos se realizan por asignaturas e incluyen un conjunto significativo de todos los materiales asociados a ella.
  + Los materiales se ofrecen de forma organizada por categorías: programa de la asignatura, lecturas obligatorias, materiales de clase, ejercicios, guía de aprendizaje,….
* El profesor o profesores garantizan que el material que publican en el OCW site es original o tiene los derechos, bien directamente por ser propietario o bien a través del tipo de licencia que los soporta, para ser reutilizados en “abierto” sin infringir los “copyrights” de otras personas.
* Son accesibles universalmente a través de la red:
  + Sin limitaciones geográficas.
  + Sin exclusión de usuario, ni necesidad de registrarse o utilizar palabras claves de acceso.
  + No exigen requisitos técnicos más allá de un navegador Web.

### BIBLIOGRAFÍA

W3C. (n.d.). *Web Services Architecture.* Retrieved 19 de 01 de 2014 from W3C: http://www.w3.org/TR/ws-arch/

Python. (n.d.). *About Python.* Retrieved 19 de 01 de 2014 from Python: http://www.python.org/about/

González, M. (n.d.). *ESTUDIO DE ARQUITECTURAS DE REDES ORIENTADAS A SERVICIO.* Retrieved 19 de 01 de 2014 from UpCommons: http://upcommons.upc.edu/pfc/bitstream/2099.1/12312/1/ESTUDIO\_DE\_ARQUITECTURAS\_DE\_REDES\_ORIENTADAS\_A\_SERVICIO.pdf

Alvarado Ruiz, P. A., Guamán Eras, D. E., & Sigcho Armijos, J. P. (n.d.). *Aplicación de tecnologías móviles para la búsqueda de recursos educativos abiertos.* Retrieved 20 de 01 de 2014 from Bibliotec UTPL: http://dspace.utpl.edu.ec/jspui/bitstream/123456789/4938/1/Pablo%20Antonio%20Alvarado%20Ruiz.pdf

UTPL. (20 de 01 de 2014). *OpenCourseWare UTPL.* From UTPL OCW: http://ocw.utpl.edu.ec/

UPM. (n.d.). *¿Qué es OCW?* Retrieved 20 de 01 de 2014 from OpenCourseWare de la Universidad Politécnica de Madrid: http://ocw.upm.es/bfque-es-ocw

WIKIPEDIA. (n.d.). *Web semántica.* Retrieved 20 de 01 de 2014 from WIKIPEDIA: http://es.wikipedia.org/wiki/Web\_Semántica

WIKIPEDIA. (n.d.). *Datos enlazados.* Retrieved 20 de 01 de 2014 from WIKIPEDIA: http://es.wikipedia.org/wiki/Datos\_enlazados

*Nltk.* (n.d.). Retrieved 21 de 01 de 2014 from NLTK: http://nltk.org/

Loper, E., & Bird, S. (n.d.). *NLTK: The Natural Language Toolkit.* Retrieved 21 de 01 de 2014 from Cornell University: http://arxiv.org/pdf/cs/0205028v1.pdf

Vallez, M., Rovira, C., Codina, L., & Pedraza, R. (n.d.). *Procedimientos para la extracción de palabras clave de páginas web basados en criterios de posicionamiento en buscadores.* Retrieved 21 de 01 de 2014 from UPF: http://www.upf.edu/hipertextnet/numero-8/extraccion\_keywords.html

*La Web Semántica y las Tecnologías del Lenguaje Humano.* (n.d.). Retrieved 23 de 01 de 2012 from e-Lis: http://eprints.rclis.org/15586/1/La%20Web%20Sem%C3%A1ntica%20y%20las%20Tecnolog%C3%ADas%20del%20Lenguaje%20Humano%20-%20Preprint.pdf