Título de la ponencia:

PROPUESTA DE ESQUEMA DE ANOTACIÓN SEMÁNTICA PARA EL SITIO WEB DE LA BIBLIOTECA DIGITAL DEL INSTITUTO CONFUCIO DE LA UNIVERSIDAD DE LA HABANA

Autora:

Dra.C. YORBELIS ROSELL LEÓN

Doctora en Ciencias Sociales por la Universidad de Granada, España. Máster en Ciencias de la Información por la Universidad de La Habana (UH), Cuba. Profesora Asistente en la carrera de Ciencias de la Información (pregrado y postgrado), Facultad de Comunicación de la UH. Directora Ejecutiva del Instituto Confucio de la UH.

Adcripción institucional:

UNIVERSIDAD DE LA HABANA

Instituto Confucio (San Nicolás #518 e/ Zanja y Dragones. Habana Vieja, La Habana)

Área de investigación:

BIBLIOTECOLOGÍA, WEB SEMÁNTICA, SISTEMAS GESTORES DE CONTENIDO

Domicilio:

Teléfono particular: (+53) 53798715

Teléfono oficina: (+53) 7861 0042 ext. 112

Correos:

yorbelisr@rect.uh.cu ; yorbelisr@gmail.com

Eje temático

HISTORIA CULTURA Y APRENDIZAJE DEL CHINO

RESUMEN

La mayoría de las bibliotecas de universidades del mundo intentan que sus productos web y la información contenida y gestionada en ellos se integren de manera eficiente. La anotación, desde el punto de vista de la web semántica, se ha convertido en un proceso que permite el marcado de puntos de acceso visibles para la recuperación de la información y para la detección de entidades o datos relevantes en el texto.

La propuesta presenta la personalización de un sistema de anotación semántica para la biblioteca digital del Instituto Confucio de la Universidad de La Habana. Esta aportaría cambios positivos a la coherencia de las políticas de difusión de la información, facilitando descripciones explícitas de los recursos de información de la biblioteca en la Web, logrando la unificación, personalización y la reutilización de la información.

PALABRAS CLAVE

BIBLIOTECA DIGITAL, ANOTACIÓN SEMÁNTICA, INSTITUTO CONFUCIO, UNIVERSIDAD DE LA HABANA

INTRODUCCIÓN

La creación de una biblioteca digital es un tema de trabajo común entre profesionales de la información, informáticos, ingenieros electrónicos, sociólogos, antropólogos, etc., todos en pos de garantizar la organización de los sistemas de información. Desde el ámbito académico, en los últimos años, su desarrollo evolutivo se adentra en el contexto digital de la web semántica y la reutilización de lenguajes universales, que permitan encontrar respuestas de forma eficiente a preguntas formuladas, desde las necesidades cognitivas de los usuarios, al generar relaciones que dotan la información de significado, entendible para los buscadores.

Este nuevo contexto ofrece a las bibliotecas con acceso desde la web, un nuevo mundo de posibilidades de interacción y consulta. Por tanto, sus servicios deben adaptarse y modificar los procesos de descripción y representación de la información que gestionan y almacenan en el entorno digital.

HIPÓTESIS

La biblioteca digital a implementar en el Instituto Confucio de la Universidad de La Habana (en lo adelante IC-UH), basa su procesamiento y gestión de información, en principios de análisis de información occidentales. La implementación de mecanismos que permitan la anotación semántica sobre los textos, imágenes y audiovisuales, por parte de los profesores chinos, contribuiría al enriquecimiento de la información contenida en los fondos digitales de la biblioteca, marcando otros puntos de acceso visibles para la recuperación de la información y para la detección de entidades o datos relevantes.

Dicha propuesta representa un salto cualitativo importante para la biblioteca en la prestación de servicios y la oferta a los estudiantes que acuden a ella, como herramienta de apoyo en el proceso de aprendizaje del chino mandarín.

PREGUNTAS DE INVESTIGACIÓN

A partir de los elementos anteriormente expuestos, en el trabajo se intenta responder a los siguientes cuestionamientos: ¿Cómo implementar una biblioteca digital en el Instituto Confucio de la Universidad de La Habana, de modo que se involucre de manera paulatina, en el entorno digital de la web semántica? ¿Cómo involucrar a los profesores de Instituto en

la construcción del fondo bibliográfico, de manera que aporten sus conocimientos sobre la enseñanza de la cultura y el idioma, como parte del procesamiento de la información?

OBJETIVO

El objetivo de la ponencia es mostrar las bases para la construcción de la biblioteca digital del Instituto Confucio que integre un sistema de anotación semántica, como herramienta de apoyo en el proceso de aprendizaje del chino mandarín.

BIBLIOTECA DIGITAL DEL CONFUCIO: PRINCIPIOS DE CONSTRUCCIÓN

En el *Digital Library Project*, se define el concepto de biblioteca digital como "el concepto de biblioteca digital no es únicamente el equivalente de repertorios digitalizados con métodos de gestión de la información. Es más bien, un entorno donde se reúnen colecciones, servicios, y personal que favorece el ciclo completo de la creación, difusión, uso y preservación de los datos, para la información y el conocimiento".

La biblioteca digital del IC-UH avistará las pautas básicas comunes a cualquier biblioteca de este tipo, con relación a los servicios previstos:

- los usuarios podrán solicitar información, independientemente de su área geográfica.
- los usuarios serán muy variados: estudiantes y profesores del propio IC-UH, comunidad del barrio chino, estudiosos de la cultura china.
- gran cantidad de información no está todavía en formato digital, pero la biblioteca digital no pretende copiar la producción impresa, sino que debe generar una nueva estructura de la información; por lo que, en lugar de digitalizarla sin más, lo mejor será recoger los metadatos de índices y catálogos de monografías y publicaciones periódicas no digitalizadas
- estará erigida sobre sistemas y plataformas convencionales que admitan la compatibilidad e integración con otros estándares.
- la biblioteca digital deberá incluir servicios de seguridad y autentificación en las transmisiones, así como el control de la propiedad intelectual.

Para insertar una biblioteca, dentro de los procesos de la Web semántica, requiere en primer lugar de la integración de una o más ontologías que permitan la gestión y almacenamiento de los datos. Partiendo de esta premisa, presentamos el diagrama conceptual de los objetos digitales modelados desde el punto de vista de las bibliotecas digitales propuesto por

Giraldo Plaza, Ruiz y Mateus (2011), para determinar cuáles serían los atributos apropiados que puedan ser modelados semánticamente (ver Figura 1)

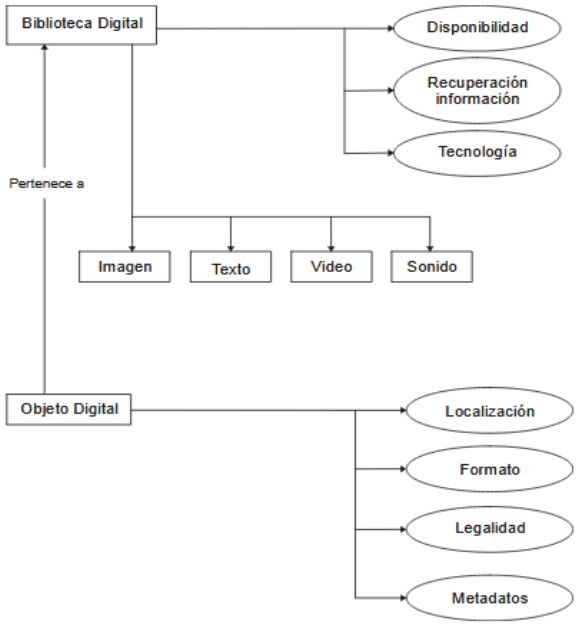

Fig. 1. Diagrama conceptual objeto digital en biblioteca digital (Fuente: Giraldo Plaza, J. E.; Ruiz, M.A; Mateus, S. P. (2011). Modelo para búsqueda y recuperación semántica en bibliotecas digitales. Revista Facultad de Ingeniería, UPTC, I Semestre 2011, vol.20, No.30, pp.31-39.)

A partir de esta modelación pueden diseñarse e implementarse una red de ontologías que respondan a la descripción y representación de las estructuras de los documentos; representación de los conceptos manejados desde los dominios temáticos y lingüísticos referentes a la cultura china; normalización y gestión de los nombres propios y autoridades en general, la gestión de anotaciones, etc.

Las ontologías integradas se implementarían para realizar consultas semánticas en un posterior desarrollo de un módulo de consulta.

Dicho lo anterior, es imprescindible la reutilización de ontologías y lenguajes ontológicos y lenguajes ontológicos que garanticen la interoperabilidad y el intercambio de información normalizada y respaldada por el proyecto W3C (por ejemplo, las disponibles en http://www.daml.org/ontologies/, *SKOS* con funciones de tesauro, *FOAF* que sirve de herramienta de control de autoridades, *BIBO* que se utiliza para describir objetos bibliográficos en la web semántica en RDF, se puede utilizar como una ontología citación, como una clasificación ontología documento, o simplemente como una forma de describir cualquier tipo de documento en RDF).

PROPUESTA DEL SISTEMA DE ANOTACIÓN SEMÁNTICA

Hemos mostrado que, con la llegada del concepto de web semántica, la revolución en el terreno de la web, los usos de ontologías y los marcadores semánticos hicieron que estas acogieran una filosofía que, si bien no desvirtuaba las formas tradicionales de navegación, dota a la web (sintáctica) de nuevas formas de razonamiento cada vez más apegadas a las necesidades de sus usuarios finales, al expresar la forma de comunicación de un dominio y sus particularidades léxicas.

La anotación es el proceso mediante el cual se describen los contenidos asociados a un sitio web, como notas, comentarios, observaciones externas de cualquier tipo o explicaciones que se pueden adjuntar al documento web o una parte seleccionada de este sin necesidad de alterar o modificar el documento original. Desde el punto de vista de la web semántica se ha convertido en un proceso que permite el marcado de puntos de acceso visibles para la recuperación de la información y para la detección de entidades o datos relevantes en el texto HTML.

Aprovechar las herramientas de construcción de la web semántica para enriquecer el funcionamiento de la biblioteca, constituye una fortaleza en el apoyo a la docencia, al mostrar una propuesta de personalización en la creación de un sistema de anotación para la biblioteca digital del IC-UH, que contribuya a una mayor efectividad en los procesos de búsqueda y recuperación de la información, que haga que los sistemas automatizados reconozcan parte de la semántica los contenidos expuestos, desde las interacciones realizadas por los usuarios. Para la propuesta se parte de sistemas ya existentes como *ANNOTEA* y *FLERSA.*

Annotea es un proyecto de la W3C, que permite hacer anotaciones en las páginas Web sin que el documento original sufra ninguna transformación. Su formato principal es RDF. Tiene un estilo de anotación semiformal, donde “las anotaciones son sentencias de texto libre sobre documentos. Estas sentencias deben tener metadatos y pueden ser clasificadas de acuerdo con un esquema RDF de complejidad arbitraria definido por el usuario” (Navarro Galindo, 2012)

El sistema de anotación propuesto requeriría considerar:

- Capacidad para la anotación semiautomática: Gestión de anotaciones semiautomáticas basadas en otros estándares de descripción que aprovechan coherentemente los metadatos de los documentos y de otras entidades no considerados como tal.
- Anotación de contenidos semánticos usuales: La anotación de contenidos que usualmente se anotan en los CMS-Semánticos, entre ellos: páginas web en formato XHTML, imágenes y elementos multimedia.
- Anotación de contenidos semánticos inusuales: La anotación de contenidos que usualmente no se anotan en los CMS-Semánticos, entre ellos: elementos connotativos, denotativos, que se van más allá de la descripción física de un recurso.
- Consistencia en la anotación: vela por el seguimiento de las anotaciones semánticas cuando existen modificaciones en la información en los documentos anotados.
- Normalización de las anotaciones: Anotaciones homogéneas y normalizadas para cada tipo de recursos en función con las necesidades de las distintas entidades de la universidad y la información que estas demandan.
- Integración de las anotaciones: La integración de las anotaciones asociadas al recurso que se describe y almacenadas en una ontología o base de conocimiento.

Se proponen tres requisitos, basado en la propuesta de Navarro Galindo (2012) que deben estar presentes en el proceso de anotación:

- El uso de una ontología tanto a nivel de infraestructura durante el proceso de creación de anotaciones semánticas, como a nivel de referencia durante el proceso de asociación de significado a los textos marcados.
- El uso de algún esquema de anotación propuesto por W3C. Para este caso particular se parte de Annotea y se enriquece en función de las necesidades de la propuesta.
- El uso de estándares de la W3C para el marcado semántico. Para este caso se implementan RDF, RDFa y OWL.

En la propuesta presentada, el framework Annotea, diseñado por la W3C y también aplicado en el proyecto FLERSA (Navarro Galindo, 2012), es la base para el manejo de la estrategia de anotación. Aunque se mantiene la taxonomía de clases de infraestructura (ver Fig. 2), al mismo se le adicionan un conjunto de clases de anotación que hacen más

potentes sus procesos de marcado semántico (Ver fig.3 y 4. Ver además https://www.w3.org/2000/10/annotation-ns#schemaClasses).

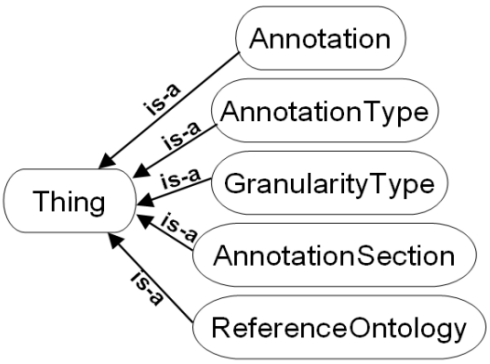

Fig. 2. Taxonomía de clases de infraestructura de Annotea (Fuente: Navarro Galindo, J. L. 2012. "FLERSA: un sistema semántico de gestión de contenido web (S-CMS)". Tesis doctoral, Universidad de Granada, España)

A Annotea se le proponen los siguientes cambios (Ver Fig.3)

Se modifican dos de las clases de Annotea:

- Clase Anotación: entre las propiedades definidas originalmente para la clase, se cambia el concepto de Autor por el concepto de Persona, pues se asume que podrían realizarse anotaciones más genéricas, no solo referidas a los autores, sino también a personajes, autoridades, investigadores, profesores, o sea, desde la perspectiva de cualquier rol desarrollado por una persona dentro de la universidad o desde una visión documental.
- Clase Sección de Anotación: se agregan dos propiedades a la clase original: Sonido y Audiovisuales, considerando que desde las universidades cubanas suelen manejarse bancos documentales de mayor riqueza.

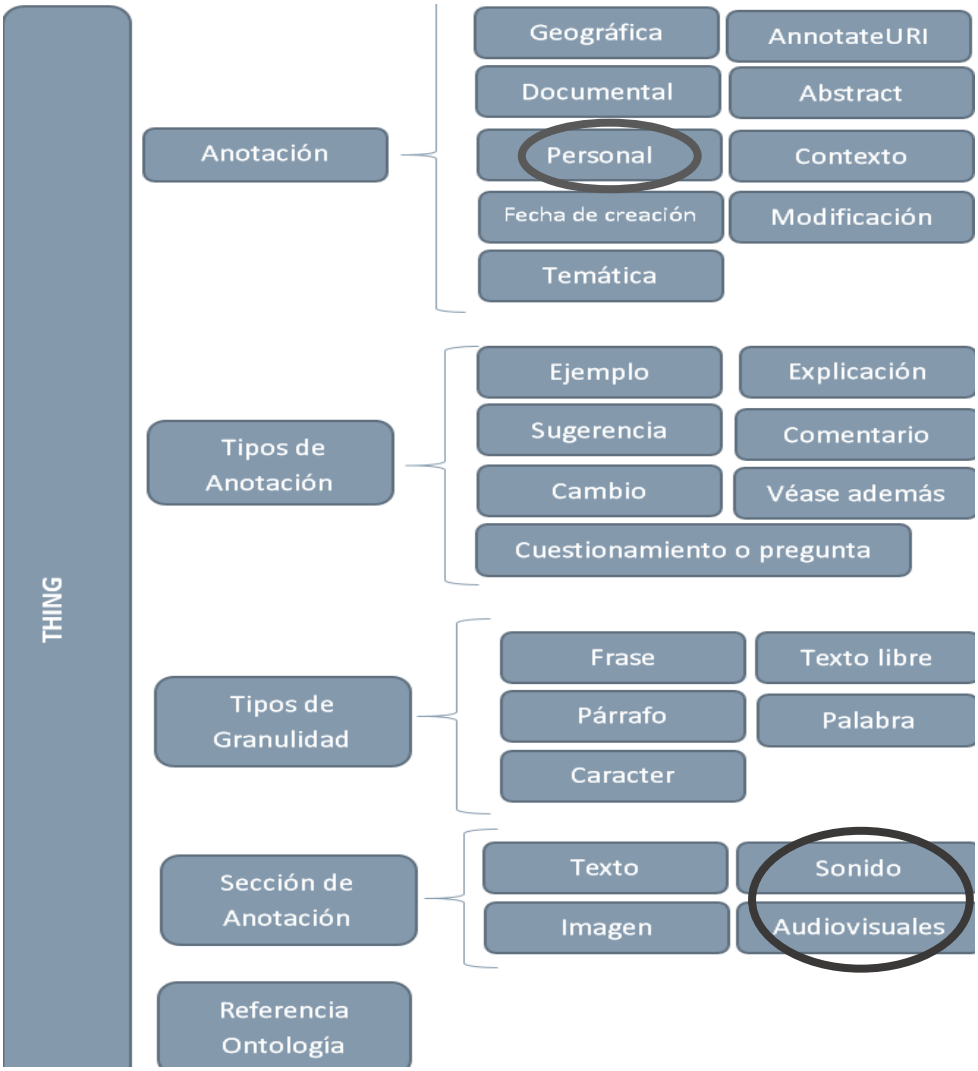

Fig. 3. Taxonomía de clases de infraestructura para las anotaciones (Fuente: elaboración propia)

Se debe diseñar e implementar una ontología para la gestión de las anotaciones, construida sobre las mismas clases de anotación de la taxonomía presentada en la figura 3, utilizando lenguajes ontológicos con un alto nivel de especificación.

A continuación, se mostrarán las clases principales propuestas para la anotación semántica:

Anotación ID: es la clase principal y dispones de las siguientes propiedades:

- **Annotate URI** (*annotateURI*): Al igual que en Annotea y en FLERSA se dedica una anotación a asociar una anotación con la página web o recurso de información sobre el que se anota. Ejemplo: http://intranet.uh.cu/
- **Área Geográfica** *(geographic)***:** Se reconoce a partir de las coordenadas establecidas en la ontología.

- **Documento** *(document):* Es un tipo de anotación que permite la selección de los datos bibliográficos de cada documento descrito en el sistema para ello se utilizan los lenguajes ontológicos Bibo y DublinCore. Esta clase de anotación genera anotaciones de tipo de frecuencia, numero, volumen, páginas, editorial.
- **Persona** *(person):* Anota no solo autores de los documentos, sino toda la información asociada la creación, modificación y manejo del documento html. Además, este tipo de anotación reconoce en el texto aquellos elementos internos del texto que se asocien a personas.
- **Resumen** *(abstract):* Sirve para anotar el texto del documento y dentro del mismo se utiliza el sistema de marcado sugerido por Leiva (2011) para desarrollar los procesos de marcación de los elementos del texto se determina organizar las unidades textuales en segmentos (Leiva Mederos, 2011). Dichos segmentos aúnan elementos de diversos órdenes como el discursivo, el sintáctico y el comunicativo, los cuales son referentes internos de las cargas semánticas y estructurales de cada oración. Las concepciones en que nos hemos basado para el diseño de las etiquetas obedece a tres criterios esenciales: la secuencia en que se etiquetan los niveles en la práctica, la simbología utilizada para declarar los elementos cohesivos del texto y la presencia en el texto de elementos geográficos, personales, documentales y de entidades. El texto se marca utilizando la ontología con sus niveles de instanciación que describe todos los procesos, maneja todas las personas de la UH.
- **Fecha de Creación** *(created):* identifica mediante una etiqueta de DublinCore denominada DublinCore: Date
- **Temática** *(subjet):* Identifica los temas principales que con carácter denotativo se encuentran en los documentos para ello cuenta con una el lenguaje ontológico *SKOS* alojado en la ontología y con las propiedades *altLabel* y *PrefLabel*, esto permite reconocer las temáticas de las que tratan los documentos mediante el análisis de texto. Generalmente las mayoría de los autores resuelven este problema usando algoritmos de agrupamiento sin explotar las bondades de la ontologías o usan las ontologías para la desambiguación de los textos, procesos de que encarece el marcado semántico, tal es el caso del algoritmo Aguirre (Aguirre, 1998) y Rigau (Rigau, 2002) el cual funciona mediante: Determinación de la totalidad de los nodos que identifican los sentidos de los

vocablos a desambiguar y los términos contextuales, lo que permite medir la densidad conceptual referente al sentido de cada término. Los elementos connotativos se marcan a partir de la propia experiencia del usuario y se aplican solo a materiales audiovisuales.

- **Contexto** *(context):* Describe la posición donde se encuentra el texto o documento audiovisual que se anota.

Se establecerán además otras clases para la anotación, como puede observarse en las figuras 4 y 5:

- **Tipos de anotación** *(type):* Asocia la anotación a un concepto dentro de la taxonomía *AnnotationType,* heredada de FLERSA, indicando el tipo de anotación. Los tipos de anotación que se pueden realizar son: ejemplo, sugerencia, cambio, cuestionamiento, explicación, comentario y véase, además.
- **Tipos de granularidad** *(granularity):* Asocia la anotación a un concepto dentro de la taxonomía *GranularityType*, heredada de FLERSA, indicando el tipo de granularidad dela anotación. Los tipos de granulidad son: caracter, palabra, frase, párrafo o texto libre. A sugerencia de Navarro Galindo (2012), el tipo de granularidad se debe asignar de forma automática dependiendo de las características del fragmento de texto que se anote, así como, no debe utilizarse en la anotación de textos multimedia (Navarro Galindo, 2012).
- **Identificación de las secciones de anotación** *(section):* se utiliza para asociar una anotación a los formatos contenidos en un recurso y sobre el cual se realiza la anotación. Por tanto, se identifica: *texto* e *imagen* al igual que en *FLRESA*, se adiciona sonido y audiovisuales, dadas las características de los recursos gestionados desde IC-UH, especialmente los de fines educativos (*learningrecourse*)
- **Referencia de ontología** (*ontologyReference*): establece la relación entre la anotación y las ontologías de referencia que se usan a modo de taxonomías. Se utilizará para asociar la anotación con el concepto del que se habla en ella.

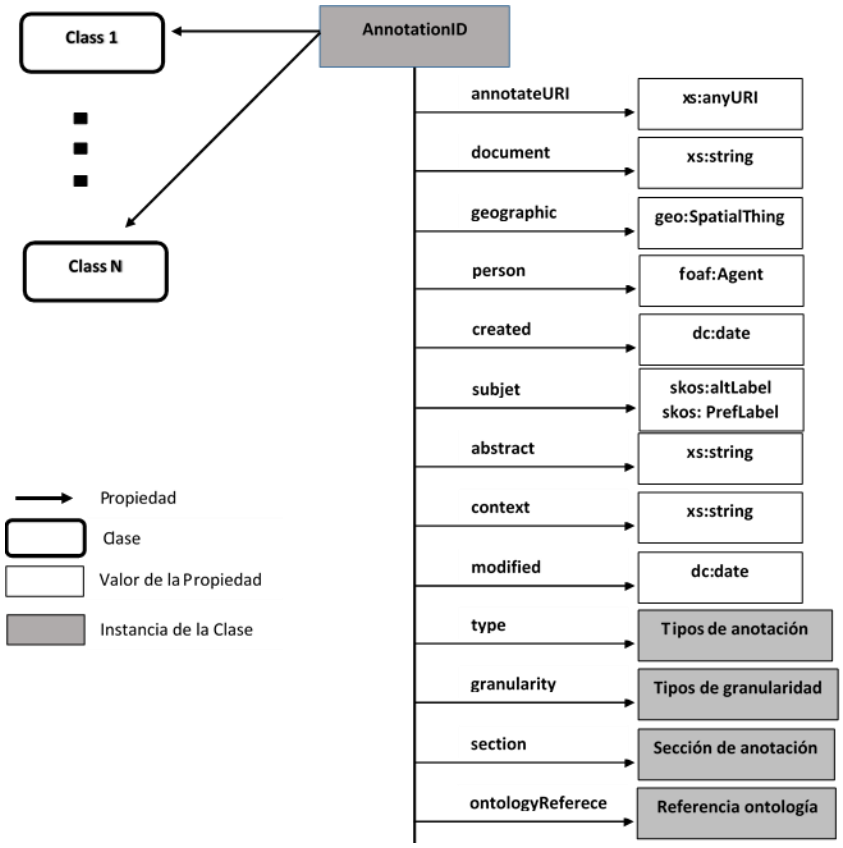

Fig. 4. Instancia de la clase AnnotationID (Fuente: elaboración propia)

Este esquema de anotación facilitará la realización de consultas en SPARQL EndPoint.

Para el modelo propuesto se trabaja sobre la base de la anotación semántica de palabras. Para ello se emplea la propuesta de Leiva (2011) para desarrollar los procesos de marcación de los elementos del texto a través de la organización de las unidades textuales en segmentos. Dichos segmentos aúnan elementos de diversos órdenes como el discursivo, el sintáctico y el comunicativo, los cuales son referentes internos de las cargas semánticas y estructurales de cada oración. El desarrollo de estas etiquetas permite la obtención de resúmenes extractos de un texto único, por tanto, en este apartado de la investigación nos centraremos en el diseño de las etiquetas necesarias para estructurar los textos, que serán la entrada del sistema.

A partir de ello, la autora para el diseño de etiquetas a partir de los siguientes criterios:

- la secuencia en que se etiquetan los niveles en la práctica,
- la simbología utilizada para declarar los elementos cohesivos del texto

- la presencia en el texto de elementos geográficos, personas y documentos.

Las marcas o etiquetas están trazadas en XML y están integradas por un tag de inicio y un tag de salida Ej:

- <***rdfs:comment*** ***xml:lang=***"*en*">
The resource in which another resource is reproduced.
</***rdfs:comment***>

Se propone, realizar el proceso mediante la herramienta XML Marker. Es una interfaz gráfica de usuario que permite a los anotadores tienen vistas simultáneas de todos los niveles de las anotaciones anteriores, mientras trabajaba en una tarea particular. Además, está equipado con instalaciones de comparación que permiten la inspección de inter-acuerdo anotador o rendimiento de la herramienta, expresada en precisión y recordar las medidas.

CONCLUSIONES

El sistema de anotación semántica para la biblioteca digital del IC-UH, convierte a la biblioteca en una herramienta de mayor potencia y eficiencia en el proceso de aprendizaje del chino mandarín, para sus usuarios potenciales.

La posibilidad de integrar a los documentos anotaciones sobre los recursos de información que la biblioteca ofrece, con las visiones y recomendaciones de los profesores del instituto, y hacer de ellas puntos de acceso para la recuperación de la información, la convierte en una herramienta de trabajo perfectamente imbricable a la docencia.

La implementación de ontologías para la gestión de datos de la biblioteca digital, fortalece su servicio y los modos de interacción con sus usuarios potenciales (profesores y estudiantes del IC-UH). Enriquece y condiciona sus catálogos para los procesos de búsqueda y recuperación de la información, al insertar la posibilidad de estrategias de pesquisa que asimilen el lenguaje natural y posibiliten inferencias semánticas a partir del reconocimiento de los usos lingüísticos del dominio particular.

Este esquema de anotación facilitará la realización de consultas a la ontología en SPARQL EndPoint.

La propuesta se basa en un modelo flexible, que puede extrapolarse a otros contextos universitarios.

Una propuesta como la que se sugiere, requiere del trabajo de un grupo interdisciplinario que integre la construcción de la biblioteca con herramientas tecnológicas desde una visión de social de las necesidades cognitivas de los estudiantes.

BIBLIOGRAFÍA

Anotaciones semánticas [Online] Available: https://www.infor.uva.es/~sblanco/Tesis/Anotaciones%20Sem%C3%A1nticas.pdf [Accessed 27 de marzo de 2017]

Giraldo Plaza, J. E.; Maryem A. Ruiz y Sandra Patricia Mateu. 2011. "Modelo para búsqueda y recuperación semántica en bibliotecas digitales". Revista Facultad de Ingeniería, UPTC, I Semestre 2011, vol.20, No.30, pp.31-39

Koivunen, M.R. 2001. Annotea: Metadata based annotation infrastructure. [Online]. Available: https://www.w3.org/Talks/2001/1102-annotea-fin/Overview.html [Accessed 12 diciembre 2017].

Koivunen, M.R. 2005. Annotea Project [Online]. Available: https://www.w3.org/2001/Annotea/ [Accessed 12 diciembre 2017].

Leiva Mederos, A. A. 2011. "Texminer: Un Modelo para el Resumen Automático y la Desambiguación de Textos Científicos en el Dominio de Ingeniería de Puertos y Costas". Tesis doctoral, Universidad de Granada, España.

Navarro Galindo, J. L. 2012. "FLERSA: un sistema semántico de gestión de contenido web (S-CMS)". Tesis doctoral, Universidad de Granada, España.

Navarro Galindo, J. L. y J. Samos. "FLERSA: Soporte a la Definición de Anotaciones y Búsquedas Semánticas en un CMS". In: Actas de las XVI Jornadas de Ingeniería del Software y Bases de Datos, 101. La Coruña, España.

Rosell León, Yorbelis. 2016. "UH-WEB: Propuesta de diseño de un CMS semántico para la Universidad de La Habana". Tesis doctoral, Universidad de Granada, España