

Raúl García Castro, Andrés García Silva

Ontology Engineering Group, Departamento de Inteligencia Artificial, Facultad de Informática, Universidad Politécnica de Madrid

<rgarcia@fi.upm.es>,
<hagarcia@delicias.dia.fi.upm.es>

1. Introducción

En los últimos años el contenido de la Web ha evolucionado, pasando de una Web formada por páginas HTML enlazadas entre sí a una Web repleta de contenidos multimedia, servicios y aplicaciones, los cuales son generados, clasificados y evaluados por los usuarios.

La búsqueda, acceso, comprensión y uso de tal cantidad de contenidos puede ser facilitada mediante la expresión de la semántica de estos recursos, para lo cual es necesario definir metadatos que los describan y asignar dichos metadatos a los recursos. Esto es lo que se conoce como el proceso de anotación semántica de recursos Web.

El concepto de "anotación semántica" podría describirse como caleidoscópico, pues adopta diferentes formas según la perspectiva desde la que se observe [1]. Explicitar el significado, es decir, la semántica de las palabras o de un documento y realizar su anotación se ha llevado a cabo de diversas maneras, dependiendo del área de investigación. En consecuencia, también son múltiples los términos acuñados en los trabajos sobre este tema para referirse a esta actividad, a veces sin diferencia aparente. En inglés, lengua en la que se publica prioritariamente, algunos de los términos utilizados son: *semantic annotation*, *semantic markup*, *semantic tagging* o *semantic labelling*.

Las distintas áreas que tratan la anotación semántica proporcionan diferentes puntos de vista sobre el proceso. En el campo de la lingüística computacional, el objetivo de la anotación es identificar y describir el significado de los textos, mediante etiquetas que corresponden a los diferentes niveles que aportan significado (morfosintáctico, semántico, discursivo, pragmático, por ejemplo). En la Web Semántica y en la Web 2.0, el objetivo de la anotación es poder incluir información no sólo de textos sino de cualquier contenido multimedia que pueda estar disponible en la Web. Mientras que en el caso de la Web Semántica se trata de anotar a partir de modelos de dominio formales (llamados ontologías), en el caso de la Web 2.0 se anota utilizando descriptores (llamados etiquetas).

Todas estas perspectivas de la anotación han de tenerse en cuenta para lograr una explotación completa del contenido de la Web, desde la agilidad del etiquetado de la Web 2.0 a la semántica formal de las ontologías. Lograr fusionar la Web Semántica y la Web 2.0 (lo

Anotación de contenidos en la Web del futuro

Resumen: en este artículo presentamos el estado y las tendencias actuales de los enfoques y herramientas de anotación de recursos Web en el contexto de las áreas de la Web Semántica, Web 2.0 y lo que se ha denominado Web 3.0, es decir, la fusión de la Web Semántica y la Web 2.0. Uno de los principales retos en la anotación de recursos Web está relacionado con la gran variedad de contenidos presentes actualmente y en el futuro, incluyendo no sólo textos sino grandes cantidades de contenidos multimedia, servicios y aplicaciones que son generados, clasificados y evaluados por los usuarios.

Palabras clave: anotación, Web Semántica, Web 2.0, Web 3.0.

Autores

Raúl García Castro es Doctor en Ciencias de la Computación e Inteligencia Artificial por la Universidad Politécnica de Madrid (UPM) y trabaja como investigador en el grupo de investigación *Ontology Engineering Group* de la Facultad de Informática de la UPM. Después de trabajar durante tres años como ingeniero de software, desde que se licenció en Informática (2003) ha estado trabajando en la UPM dentro del *Ontology Engineering Group*, donde defendió su tesis doctoral en 2008. Ha trabajado en varios proyectos de investigación nacionales y europeos centrándose en la evaluación y el *benchmarking* de tecnologías de la Web Semántica. Participa habitualmente en los comités de programa de las conferencias y talleres más relevantes del área y ha organizado talleres y conferencias internacionales.

Andrés García Silva es candidato a doctor en Inteligencia Artificial por la Universidad Politécnica de Madrid (UPM). Previamente obtuvo los títulos de Máster de investigación en Inteligencia Artificial por la UPM (2008), de Especialista en Administración de Empresas por la Universidad ICESI de Cali (Colombia, 2005), de Ingeniero de Sistemas por la Universidad del Valle de Cali (2001) y de Profesional certificado en Oracle por Oracle Corporation (2003). Tiene especial interés en la Web Semántica, la Web 3.0 y sus áreas relacionadas, incluyendo el desarrollo de ontologías. Es miembro del *Ontology Engineering Group* perteneciente al Laboratorio de Inteligencia Artificial de la Facultad de Informática de la UPM. Además ha participado en los proyectos de investigación europeos NeOn y SEEMP, en el proyecto español Geobuddies y en el proyecto CENIT España Virtual.

que se ha denominado la Web 3.0) permitirá obtener lo mejor de los dos mundos e integrar las etiquetas de la Web 2.0 con la semántica formal de las ontologías.

En este documento se resumen los distintos enfoques adoptados en la anotación en los campos de la Web Semántica y la Web 2.0, y la Web 3.0, y se identifican tendencias actuales de trabajo.

2. Anotación en la Web Semántica

En la comunidad informática el término *marcado semántico* (*semantic markup*) se empleó inicialmente para denotar la descripción de los contenidos de una página Web adjuntando a ciertos elementos del texto una etiqueta que identifica de alguna manera su referente en el mundo real (en este caso, en términos computacionales, se emplea lo que se conoce como *ontologías*). Para codificar el significado se utilizaban los denominados lenguajes de marcado, tales como XML [2], RDF(S) [3] u OWL [4].

Según Shea [5], el marcado de una página

Web puede ser de dos tipos, marcado para la presentación (*presentation markup*) y marcado semántico (*semantic markup*). El *marcado para la presentación* consiste en etiquetar una página Web con información sobre las características formales del documento mediante HTML (es decir, cómo debe mostrarse al lector, con negrita, mayúsculas, colores, etc.). El *marcado semántico*, en cambio, incluiría información sobre el contenido de la página Web en cuestión (documento, imagen, URL, etc.) y estaría implementado en XML o en cualquier otro lenguaje de marcado basado en él.

En cualquier caso, la preparación de documentos para su publicación en la Web Semántica mediante anotaciones que expliquen su contenido y que sean procesables por un ordenador conlleva al menos tres posibles fases: (1) la elección del lenguaje que se utilizará; es decir, ha de ser uno de los lenguajes basados en XML y desarrollados para la Web Semántica, como RDF, RDF(S) u OWL, entre otros; (2) la adición, en el lenguaje elegido, de comentarios o anotaciones semánticas (metadatos sobre los documen-



"Una buena parte de la investigación en el campo de la Web Semántica se dirige hacia la anotación de datos multilingües"



tos o páginas Web que aportan cierto significado) a los recursos que desean publicarse en la Web Semántica; y (3) el razonamiento con estos metadatos semánticos. Una condición fundamental para que estos metadatos tengan valor y permitan el razonamiento es su especificación de forma consensuada para que pueda ser compartida por los distintos agentes participantes. Precisamente para alcanzar este consenso sobre el significado nacen las ontologías. Una *ontología* se define como una "especificación formal y explícita de una conceptualización compartida" [6]. De forma informal, se puede decir que una ontología es un modelo de conocimientos consensuado en un determinado dominio, que proporciona un vocabulario compartido y controlado, que posibilita la comunicación entre personas y aplicaciones.

Las ontologías aparecen como la espina dorsal de la anotación de documentos en las aplicaciones previas a la Web Semántica, como el proyecto SHOE, la iniciativa (KA)² y el proyecto Planet-Onto, entre otros. Con la aparición de la Web Semántica, la anotación basada en ontologías ha sido el foco de numerosos proyectos y aplicaciones, dado que la existencia de contenido anotado es uno de los retos más importantes para conseguir hacerla realidad. Entre estos proyectos y aplicaciones se pueden citar los proyectos europeos Esperonto¹ y AceMedia², la Red de Excelencia Europea SCHEMA³, o el proyecto estadounidense MindSwap⁴. Todos ellos comparten la premisa de proporcionar herramientas y marcos de trabajo para anotar distintos tipos de contenidos (HTML, bases de datos, documentos multimedia, etc.) con distintos niveles de detalle. El estado de la cuestión sobre anotación semántica ha sido analizado en [7] y [8]⁵.

Existen diversos grados de formalidad en los que se pueden realizar las anotaciones de un documento. En líneas generales se puede distinguir entre anotaciones que utilizan vocabularios como *Dublin Core* (que describen en líneas generales el documento sin entrar en detalles sobre su contenido concreto), anotaciones basadas en tesauros y listas de términos (que describen el contenido de los documentos relacionándolo con términos que aparecen en dichos vocabularios), y anotaciones basadas en ontologías (que se centran en la creación de instancias de clases y relaciones de ontologías formalmente especificadas

en algún lenguaje como RDF Schema u OWL y que, por tanto, son las de mayor precisión).

Una clasificación detallada de los distintos tipos de anotación que pueden aplicarse a páginas de la Web Semántica puede encontrarse en [9]. Por su parte, en [10] se incluye una discusión acerca del grado de formalización que alcanzan los distintos tipos de anotación posibles.

2.1. Anotación multimedia

La anotación de documentos multimedia todavía sigue siendo un área en el que se está trabajando intensamente, y existen grandes proyectos en ejecución donde uno de sus objetivos principales es el de proporcionar herramientas que permitan realizar estos tipos de anotaciones, tal como X-Media⁶ [11], y otros proyectos pasados en los que también se ha trabajado en esta área, tales como Acemia⁷ [12], MUMIS⁸ o Esperonto [13]. También hay un grupo activo asociado al consorcio W3C que trabaja proporcionando recomendaciones⁹.

Algunas de las propuestas consisten en la anotación de los textos que acompañan a una determinada imagen en un documento, dado que normalmente dichos textos tratarán sobre la imagen en cuestión y pueden proporcionar información interesante sobre la misma. Sin embargo, este tipo de anotación de archivos multimedia no proporciona suficiente calidad, dado que no siempre el texto da suficiente información sobre los archivos adjuntos.

En otro sentido, uno de los principales problemas que debe abordarse al realizar anotación de archivos multimedia, especialmente en el caso de imágenes y vídeos, es lo que se conoce como la *brecha semántica*. Por un lado, las técnicas de análisis de imágenes permiten obtener una gran cantidad de descriptores de una imagen, basadas en las características de sus píxeles, y permiten incluso identificar objetos y su posición con respecto a la imagen. Sin embargo, aún es difícil pasar de ahí a la anotación de los objetos concretos que se encuentran en la imagen y, sobre todo, la relación entre ellos (por ejemplo, "esta foto trata sobre un niño que está en la playa"). Este es precisamente el foco de los proyectos AceMedia o X-Media, entre otros.

2.2. Tecnología actual de anotación

De manera muy básica, las herramientas de

anotación basadas en ontologías permiten poblar ontologías con instancias procedentes de distintos tipos de recursos. Debido a la heterogeneidad de estos recursos, estas herramientas son muy heterogéneas. Algunas cubren todo el ciclo de vida de la anotación y otras sólo parte del mismo y, aunque gran parte de ellas trabajan sobre documentos de texto, algunas ofrecen la posibilidad de anotar contenido de la Web o recursos multimedia.

La mayor parte de las herramientas de anotación permiten la anotación manual o semiautomática de contenidos, estas últimas utilizando técnicas basadas en extracción de información o aprendizaje automático. La anotación manual de documentos es una tarea costosa y que puede llevar a gran número de errores, como se ha demostrado en la mayor parte de las iniciativas de la Web Semántica y de las anteriores a ésta. Por tanto, el esfuerzo en anotación debería centrarse en la automatización de esta tarea, dejando a los usuarios la tarea de supervisión de las anotaciones propuestas. Para ello, la calidad de la anotación automática debe también mejorar, con el objetivo de hacer más productivo todo este proceso.

Además de tratar con la calidad de las anotaciones generadas, es importante que las herramientas de anotación gestionen adecuadamente las inconsistencias que se pueden producir en los datos anotados de manera distribuida, el ciclo de vida de las anotaciones y de sus vocabularios relacionados (por ejemplo, la gestión de la evolución de los vocabularios en los cuales las anotaciones están basados), la existencia de sistemas de gestión de anotaciones para realizar consultas, almacenamiento, razonamiento, etc. Todo esto son líneas que se están tratando actualmente.

2.3. Tendencias

Hoy en día, una buena parte de la investigación en el campo de la Web Semántica se dirige hacia la anotación de datos multilingües, con el fin de que las búsquedas de información recuperen los documentos relevantes sea cual sea la lengua en la que estén escritos. También se trabaja en el tratamiento de la calidad de las anotaciones generadas.

Por otro lado, en el área de anotación semántica de documentos multimedia todavía queda mucho trabajo por realizar, y la mayor parte de los esfuerzos se centran en el proble-

ma de la brecha semántica, que se puede definir como el problema de pasar de las representaciones basadas en análisis de imágenes y vídeos a representaciones semánticas de las mismas, en las que se dice que un objeto concreto es por ej. una mesa o una silla, y que dicho objeto está relacionado con otro de los que aparecen en la imagen o vídeo de alguna manera (está encima de él, una persona está sentada en una silla, etc.).

3. Anotación en la Web 2.0

En los últimos años se ha pasado de una Web donde el contenido era generado por los creadores de los sitios Web a una Internet más abierta y social donde los usuarios ya no son sólo consumidores de información sino también creadores. Esta nueva etapa, conocida como la Web 2.0¹⁰, trajo consigo una variedad de aplicaciones como wikis, blogs, redes sociales y sitios Web para compartir direcciones en Internet, fotos, vídeos o música. Toda esta nueva gama de aplicaciones conlleva nuevos contenidos para los cuales los métodos tradicionales de anotación y recuperación de información no son suficientes. Surge entonces la anotación mediante la asignación de palabras clave a recursos, los cuales pueden posteriormente ser recuperados utilizando dichas palabras clave. Estas palabras clave se conocen como etiquetas y el proceso de asignación de etiquetas se denomina etiquetado.

Tener el contenido de la Web etiquetado permite clasificarlo, adaptarlo y relacionarlo con otros contenidos. Cualquier usuario puede etiquetar el contenido de la Web asignando las etiquetas que quiera y cuantas quiera, lo que resulta más fácil y más flexible que encajar la información en carpetas o categorías ya establecidas. A la colección de etiquetas generada por la anotación colaborativa de recursos se le denomina *folksonomía*. La forma más utilizada para visualizar las etiquetas de una *folksonomía* es por medio de las nubes de etiquetas, donde las etiquetas más recientes y más usadas se presentan en diferentes tamaños de fuente, donde una mayor fuente indica una mayor frecuencia de uso.

La anotación colaborativa ha tenido éxito ya que los usuarios no necesitan habilidades específicas para llevarla a cabo. Los usuarios no tienen ninguna restricción sobre las etiquetas que pueden asignar. Sin embargo, esta libertad en el uso de las etiquetas y la carencia de relaciones explícitas entre etiquetas hace que las *folksonomías* presenten problemas que afectan el proceso de recuperación de información. Por ejemplo, debido al uso de etiquetas polisémicas un usuario puede recuperar información que no es de su interés. O, debido a variaciones morfológicas de una palabra o al uso de sinónimos o de palabras compuestas, el usuario puede no encontrar la información que requiere. Es por estos pro-

blemas que las *folksonomías* hoy en día se consideran más adecuadas para la exploración de la información con el objetivo de encontrar nuevos contenidos que para la búsqueda de información en sí.

Por otra parte, los sitios Web que utilizan etiquetado colaborativo no siguen un estándar para el almacenamiento ni para la publicación de las *folksonomías*. Esto hace que estos sitios sean islas donde un usuario tiene que crear un conjunto de etiquetas para anotar sus fotos y otro independiente para anotar su blog, siendo probable que exista una relación entre las fotos que ha cargado y el blog que está escribiendo, como podría ser el caso de una descripción de un viaje con sus respectivas fotos. Permitir el intercambio de información de la *folksonomía* entre diferentes redes sociales es uno de los retos que afronta la Web 2.0 actual.

3.1. Modalidades de etiquetado

El etiquetado se puede analizar desde múltiples dimensiones como las que se han identificado en [14] y que se refieren al tipo de soporte a la anotación que ofrece la aplicación de etiquetado colaborativo, al modelo de agregación que utiliza, y a quien puede realizar la anotación.

■ **Soporte a la anotación.** Este factor distingue tres categorías: a) anotación ciega, cuando el usuario no ve las etiquetas que otros usuarios han asignado al mismo recurso; b) anotación visible, cuando el usuario ve las etiquetas que otros han asignado; y c) anotación sugerida, cuando las etiquetas se sugieren en base a etiquetas existentes del mismo usuario o a etiquetas asignadas al mismo recurso por otros usuarios, entre otros casos.

■ **Modelo de agregación.** El sistema permite o no asignar múltiples veces la misma etiqueta a un mismo recurso por diferentes usuarios. Se distingue entre el modelo de bolsa que admite etiquetas duplicadas como en Delicious, y el modelo de conjunto en el que no se admiten etiquetas duplicadas como en Flickr.

■ **Derechos de anotación.** Distingue quién etiqueta el recurso: sólo el dueño del recurso (auto-etiquetado), a quien él le haya dado permiso implícitamente por ser miembro de su red social o explícitamente (con permiso), o cualquier usuario puede etiquetar cualquier cosa (libre para todos).

Algunos sistemas no solicitan las etiquetas por medio de una interfaz de usuario convencional sino que exigen que se incluyan como elementos de código especiales que luego son procesados automáticamente. Por ejemplo, si un usuario quiere que las etiquetas asociadas a una entrada sean recopiladas junto a ésta, debe escribir las anotaciones en XML dentro del código de la entrada siguiendo un formato predefinido. Así, cuando Technorati encuentra la entrada, la analiza y, si encuentra las etiquetas

dentro del código, las carga en su sistema y las pone a disposición de los usuarios.

3.2. Recomendación de etiquetas

Los sistemas de anotación colaborativa generalmente incluyen algún tipo de mecanismo de recomendación que ayuda a los usuarios a encontrar las etiquetas más apropiadas para anotar un recurso. Etiquetar un recurso con un buen grupo de etiquetas ayuda en el descubrimiento y la recuperación de información. Una etiqueta específica ayuda a identificar fácilmente un recurso pero puede ser menos útil para que otros usuarios descubran nuevos objetos. Por el contrario, una etiqueta genérica es útil para descubrir recursos, pero no es efectiva para recuperar recursos específicos. Además, la recomendación de etiquetas es importante porque ayuda a estabilizar rápidamente el vocabulario utilizado dentro de la *folksonomía*.

ZoneTag [15] es una aplicación móvil para la captura y anotación de fotos que son cargadas a Flickr. Esta aplicación recomienda etiquetas teniendo en cuenta el contexto de la foto. El contexto para ZoneTag se compone de la localización y del tiempo. Las etiquetas recomendadas se agrupan en tres categorías que reflejan su origen: (1) etiquetas locales, que son creadas por el usuario, por miembros de su red social o por otros en su localización actual, (2) etiquetas recientes, son las usadas por el usuario las últimas veinticuatro horas y (3) todas las etiquetas. Además, ZoneTag incorpora etiquetas de otras fuentes, como lugares y eventos, basándose en la localización física.

En Flickr existe la opción *Elige de tus etiquetas* que despliega en orden alfabético las etiquetas que ya han sido utilizadas por el usuario para anotar las fotos. Para dotar a Flickr de una recomendación de etiquetas más sofisticada en [16] se presenta un enfoque basado en la co-ocurrencia de etiquetas. El sistema genera una lista de etiquetas candidatas que co-ocuran con las etiquetas que el usuario ha asignado inicialmente, luego estas etiquetas son agrupadas y ordenadas.

Además, para recomendar etiquetas se han utilizado técnicas de aprendizaje de máquina y filtrado colaborativo. En [17] se han usado técnicas de aprendizaje de máquina y procesamiento de lenguaje natural con el objetivo de encontrar páginas previamente anotadas por el usuario similares a la página que se desea anotar y luego sugerir las etiquetas de las páginas similares. El filtrado colaborativo también se ha utilizado para recomendación de etiquetas. En [18] se propone adaptar la estructura tripartita de las *folksonomías* a un espacio bidimensional y así poder utilizar las técnicas de filtrado colaborativo tradicionales.

3.3. Tendencias

Actualmente las aplicaciones de la Web 2.0

han empezado a generar funcionalidades que permiten reducir la ambigüedad de las etiquetas y proveer de estructura a la información que codifican las etiquetas. Para reducir la ambigüedad de las etiquetas que pueden tener más de un significado, Delicious permite que los usuarios asignen una descripción a cada etiqueta. De la misma manera, se permite crear grupos de etiquetas con el fin de organizarlas mejor.

Por otra parte, los usuarios de Flickr cada vez empiezan a utilizar más las llamadas etiquetas de maquina como "flora:tree=coniferous", "medium:paint=oil" y "flickr:user=straup". Estas etiquetas constan de un espacio de nombres, seguido de un predicado y un valor. El espacio de nombres define una clase o categoría a la cual la etiqueta pertenece y el predicado es una propiedad del espacio de nombres a la cual se le asigna el valor. Es posible realizar consultas como "quiero recuperar todas las fotos que tengan un título", utilizando toda una expresión como "machine_tags" => "title=".

4. Anotación en la Web 3.0

En la Web 3.0 se utilizan combinaciones de técnicas de anotación de la Web Semántica y de la Web 2.0 con el objetivo de minimizar problemas de ésta como son la ambigüedad y la falta de relaciones explícitas entre las etiquetas que forman parte de una *folksonomía*. Además de ofrecer portabilidad a los datos de la Web 2.0, las tecnologías de la Web Semántica pueden mejorar el proceso de anotación ofreciendo etiquetas más acordes al dominio y al contenido de los recursos, utilizando ontologías que modelen el conocimiento del dominio en cuestión. Por último, formalizar la información de etiquetado mediante lenguajes de la Web Semántica permite que dicha información sea procesable automáticamente y que esté disponible en buscadores semánticos como Swoogle y Watson.

Los lenguajes de la Web Semántica como RDF son útiles porque favorecen el compartir información y su procesamiento automático, dos características de las que carece actualmente la Web 2.0. Además, las ontologías permiten hacer explícita y formal la semántica de las etiquetas que forman parte de las *folksonomías*, con lo cual se mejoran los procesos de recuperación de información. De la misma manera, las ontologías permiten incluir información semántica de los usuarios o la descripción de los recursos a los procesos actuales de filtrado colaborativo, lo que permite recomendaciones más precisas.

En [19] se afirma que las ontologías para la anotación son útiles porque contribuyen a:

- Sofisticar la representación del conocimiento. Una ontología de anotación puede representar robustamente entidades y relaciones que dan forma a las actividades de anotación, estableciendo explícitamente la estruc-

tura de conocimiento de los datos de la anotación y facilitando el enlace de datos de la anotación en la Web.

- Facilitar el intercambio de conocimiento. Las ontologías habilitan el intercambio de conocimiento entre diferentes usuarios y aplicaciones al proveer construcciones reutilizables. Así, una ontología de anotación puede ser compartida y utilizada para distintas actividades de anotación en diferentes plataformas.
- Hacer la información procesable automáticamente. Las ontologías y la Web Semántica en general exponen el conocimiento humano a las maquinas de tal manera que pueda ser enlazada e integrada con otras fuentes de información automáticamente.

4.1. Ontologías para etiquetados colaborativos

Algunos modelos para representar, identificar y formalizar una conceptualización de la actividad de etiquetado a nivel semántico han sido últimamente propuestos. Describimos seguidamente las ontologías más relevantes para modelar el proceso de etiquetado colaborativo, descentralizado y espontáneo.

TagOntology [20] es una ontología para el proceso de anotación que incluye conceptos para modelar el recurso a etiquetar, la etiqueta que se utiliza, y quien realiza el etiquetado. Además se incluye el concepto de polaridad con el que se puede valorar positiva o negativamente la anotación. La Tag Ontology¹¹ de Newman tiene una estructura similar a la anterior. En particular la ontología de Newman define el anotador como un agente FOAF (*Friend Of A Friend*).

La ontología SCOT¹² *Social Semantic Cloud of Tags* adiciona información a las etiquetas incluyendo su frecuencia, la frecuencia de coocurrencia con otras etiquetas e información de equivalencia con otras etiquetas. También incluye el concepto *TagCloud* (nube de etiquetas) para definir conjuntos de etiquetas.

La ontología MOAT¹³ es la única que permite definir el significado de la etiqueta. El significado se define por medio de la asignación de URIs de recursos de la Web Semántica (tales como URIs de DBPedia o GeoNames).

Por último, en [21] se ha evaluado la posibilidad de formalizar el proceso de anotación utilizando SKOS¹⁴. SKOS es un vocabulario útil para expresar sistemas de representación de conocimiento como tesauros o taxonomías sencillas. Modelar *folksonomías* por medio de SKOS presenta varias dificultades: Modelar las etiquetas de una *folksonomía* mediante un *skos:Concept* no es semánticamente correcto porque la clase *skos:Concept* debe referirse a una idea o noción abstracta, es decir, a un concepto bien definido formalmente con un significado de consenso para todos sus usuarios. Esta clase

permite declarar que un recurso es conceptual. Un símbolo de una *folksonomía* sin embargo, no llega a serlo porque cada usuario le asigna uno o varios significados.

Hay una propuesta para resolver el problema del modelado de propiedades léxicas en SKOS que garantiza la compatibilidad con OWL-DL. Se trata de extender SKOS con la posibilidad de definir etiquetas léxicas para un *skos:Concept* o un recurso. Se adiciona la clase *skos:Label* (y sus correspondientes propiedades *skos:prefLabelResource*, *skos:altLabelResource*, *skos:hiddenLabelResource*) que se puede utilizar como vínculo entre el concepto y su etiqueta. Además se introduce una nueva propiedad *skos:labelRelation* para expresar relaciones entre etiquetas, de manera que se pueda especializar para definir relaciones particulares.

4.2. Tendencias

La investigación actual y futura en temas de la Web 3.0 incluye, entre otros temas, los derivados de la carencia de una semántica explícita en las *folksonomías*. Se han presentado algunos enfoques que buscan identificar automáticamente la semántica oculta en el conjunto de etiquetas de una *folksonomía* y, posteriormente, formalizar ésta información por medio de ontologías. Estos enfoques utilizan técnicas estadísticas o de minería de datos para encontrar grupos de etiquetas, o técnicas lingüísticas y de gestión de conocimiento cuando se trata de identificar explícita y formalmente el significado de cada etiqueta. A partir de los primeros resultados obtenidos han surgido nuevos problemas como por ejemplo el de la evolución conjunta de la *folksonomía* y la ontología resultante del proceso de identificación semántica, y también el sentido contrario, es decir, si la ontología resultante cambia como se vería afectada la *folksonomía*.

5. Conclusiones

Como se muestra en este artículo, la investigación en el campo de la Web Semántica se dirige hacia la anotación de datos multimedia (textos, imágenes, vídeo, etc.), de forma que se pueda proporcionar información acerca de distintos contenidos de forma integrada, así como hacia la anotación multilingüe, con el fin de que las búsquedas de información recuperen los documentos pertinentes sea cual sea la lengua en la que estén escritos.

Por otro lado, la investigación en el campo de la Web 2.0 se dirige a la integración de datos de distintas redes sociales mediante la identificación de relaciones entre *folksonomías*, para permitir un acceso integrado a los datos a lo largo de las distintas aplicaciones, y a la anotación de etiquetas para dotarlas de significado y poder procesarlas automáticamente.

No obstante, aún existen numerosos problemas abiertos que deberán ser tratados en mayor o menor medida. Uno de estos problemas es

debido a la integración de datos de distintas fuentes. En el entorno abierto de la Web, esta integración provocará inconsistencias e información incompleta que habrá que manejar, tanto a nivel de los datos como de los modelos de dichos datos (ontologías o *folksonomías*).

Por otro lado, el incremento de información multimedia en la Web requiere dar soluciones para representar la información en los distintos documentos multimedia y para relacionar contenidos multimedia de distinta naturaleza (porej., relacionar las imágenes y vídeos que aparecen en un documento y su texto). Estas soluciones son abordadas parcialmente en la Web Semántica y la Web 2.0, pero una solución general del problema requiere una aproximación híbrida.

Otra situación relevante es la del manejo del ciclo de vida de las anotaciones (evolución de las anotaciones, gestión de cambios, diferentes versiones, etc.) que en muchos casos está relacionada con la procedencia de dichas anotaciones y de los datos que anotan.

Además, hay que destacar que las áreas de investigación tratadas en este documento son muy dinámicas y se espera que su evolución traiga nuevos avances tecnológicos. Actualmente las herramientas de anotación se dividen en herramientas orientadas a usuarios expertos, no usables por un público en general, y herramientas de dominio específico, difíciles de integrar con otras.

En el futuro se espera pasar de las herramientas actuales de anotación, manuales o semiautomáticas, a herramientas de anotación automáticas en las que el papel del usuario sea el de supervisar y evaluar la calidad de las anotaciones. Por otro lado, se espera en el futuro que las herramientas de anotación estén parcial o totalmente integradas con herramientas y sistemas de uso general, tales como herramientas de comunicación personal (chats, mensajería instantánea, correo electrónico, etc.), herramientas de escritorio, herramientas ofimáticas, etc.

Agradecimientos

Este trabajo ha sido parcialmente subvencionado por el proyecto CENIT "España Virtual", subvencionado por el CDTI dentro del programa Ingenio 2010. Muchas gracias a los siguientes miembros del *Ontology Engineering Group* por sus contribuciones a este artículo: Guadalupe Aguado de Cea, Inmaculada Álvarez de Mon Rego, Óscar Corcho y Antonio Pareja Lora.

Referencias

- [1] G. Aguado de Cea, I. Álvarez de Mon, A. Pareja-Lora. *Una visión interdisciplinar de la anotación semántica*. Amparo Alcina (ed.). Terminología y Sociedad del conocimiento. Peter Lang, Berna 2008.
- [2] T. Bray, J. Paoli, C. Sperberg. *Extensible Markup Language (XML) 1.0*. W3C Recommendation, 1998. <<http://www.w3.org/TR/REC-xml>>
- [3] F. Manola, E. Miller. *RDF Primer W3C Recommendation*. 10 Febrero 2004. <<http://www.w3.org/TR/rdf-primer/>>
- [4] D.L. McGuinness, F. van Harmelen. *OWL: Web Ontology Language Overview*. (W3C Recommendation, 10 February 2004). <<http://www.w3.org/TR/owl-features/>>
- [5] D. Shea. *Who Cares about Semantics Anyway?* <http://www.mezzoblue.com/archives/2005/05/30/who_cares_ab/> [Accedido en octubre, 2008]
- [6] R. Studer, VR. Benjamins, D. Fensel. *Knowledge Engineering: Principles and Methods*. *IEEE Transactions on Data and Knowledge Engineering* 25(1-2): pp.161-197. 1998.
- [7] O. Corcho. *Ontology based document annotation: trends and open research problems*. *International Journal of Metadata, Semantics and Ontologies* 1(1), 2005.
- [8] S. Handschuh, S. Staab. *Annotation for the Semantic Web*. IOS Press, 2003.
- [9] S. Bechhofer, L. Carr, C. Goble, S. Kampa, T. Miles-Board. *The Semantics of Semantic Annotation*. *ODBASE: First International Conference on Ontologies, Databases, and Applications of Semantics for Large Scale Information Systems, 2002*. Irvine, California.
- [10] M. Uschold. *Where are the Semantics in the Semantic Web?* *AI Magazine*, 24 (3), 2003: 25-36. Menlo Park (California): American Association for Artificial Intelligence (AAAI).
- [11] Kosmas Petridis, Dionysios Anastasopoulos, Carsten Saathoff, Norman Timmermann, Yiannis Kompatsiaris, Steffen Staab. *M-OntoMat-Annotizer: Image Annotation Linking Ontologies and Multimedia Low-Level Features* *KES 2006*.
- [12] Ajay Chakravarthy, Vitaveska Lanfranchi, Daniela Petrelli, Fabio Ciravegna. *Semantic Enrichment of Multimedia Document*. *First International Workshop on Semantic Web Annotations for Multimedia (SWAMM)*. 2006.
- [13] T. Declerck, J. Contreras, O. Corcho, C. Crispi. *Text-based Semantic Annotation Service for Multimedia Content in the Esperanto Project*. *European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies (EWIMT2004)*. London, United Kingdom.
- [14] Cameron Marlow, Mor Naaman, Danah Boyd, Marc Davis. *HT06, tagging paper, taxonomy, Flickr, academic article, to read*. *HYPERTEXT '06: 17th conference on Hypertext and hypermedia*, pp. 31-40. ACM Press.
- [15] Morgan Ames, Mor Naaman. *Why We Tag: Motivations for Annotation in Mobile and Online Media*. *Actas de la CHI 2007*, San Jose, CA, USA.
- [16] Börkur Sigurbjörnsson, Roelof van Zwol. *Flickr tag recommendation based on collective knowledge*. *17th International Conference on World Wide Web (WWW 2008)* pp. 327-336.
- [17] P. Basile, D. Gendarmi, F. Lanubile, G. Semeraro. *Recommending Smart Tags in a Social Bookmarking System*. *Proc. of International Workshop on Bridging the Gap between Semantic Web and Web 2.0. Actas de la 4th European Semantic Web Conference (ESWC'07)*.
- [18] Robert Jäschke, Leandro Marinho, Andreas Hotho, Lars Schmidt-Thieme, Gerd Stumme. *Tag Recommendations in Folksonomies*. *Knowledge Discovery in Databases: PKDD 2007*.
- [19] H.L. Kim, S. Scerri, J.G. Breslin, S. Decker, H.G. Kim. *The State of the Art in Tag Ontologies: A Semantic Model for Tagging and Folksonomies*. *The 8th International Conference on Dublin Core and Metadata Applications*. Berlin, Germany, Septiembre 2008.
- [20] Thomas Gruber. *TagOntology, a way to agree on the semantics of tagging data*, 2005.
- [21] Emilio Rubiera, Sergio Fernández. *Morfeo Project, Entregable 5.3.1*. Integración de folksonomías y ontologías, 2007.

Notas

- 1 <<http://www.esperanto.net/>>
- 2 <<http://www.acemedia.org/>>
- 3 <<http://www.schema-ist.org/>>
- 4 <<http://www.mindswap.org/>>
- 5 <<http://annotation.semanticweb.org/>>
- 6 <<http://www.x-media-project.org/>>
- 7 <<http://www.acemedia.org/>>
- 8 <<http://lands.let.kun.nl/TSPublic/MUMIS/>>
- 9 <<http://www.w3.org/TR/swbp-image-annotation/>>
- 10 <<http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html>>
- 11 <<http://www.holygoat.co.uk/projects/tags/>>
- 12 <<http://scot-project.org/>>
- 13 <<http://moat-project.org/>>
- 14 <<http://www.w3.org/2004/02/skos/>>