Actualización del concurso simultáneo en el uso del lenguaje libre y del controlado: folksonomías y taxonomías.

José Antonio Moreiro González, Jorge Morato Lara, Sonia Sánchez-Cuadrado, Vicente Palacios.
Universidad Carlos III de Madrid, España

Introducción

Tradicionalmente los sistemas de información han venido empleando diversos lenguajes que posibilitan la mejora en la recuperación de los documentos. Los soportes actuales no han podido decidir cual de ellos es el más idóneo debido a que todos presentan ventajas e inconvenientes de uso. Estos lenguajes ofrecen variaciones en su grado de estructuración, desde los lenguajes libres a los más controlados y formalizados, como se muestra en el siguiente esquema:

- Palabras-clave independientes que se usan en indización libre tanto por extracción como por asignación, uno de cuyos tipos son las folksonomías.
- Listas de palabras, como los glosarios, listas de nombres, diccionarios, donde ahora se incluyen también los anillos semánticos (*hypernymy tree* en *wordnet*) (Gómez, 2004).
- Facetas, categorizaciones y clasificaciones: lenguajes propiamente taxonómicos o con esquemas categorizadores (Zeng, 2004) (Z 39.19-2005).
- Grupos de relaciones, basados en relaciones entre términos y

conceptos, de estructura más compleja, entre los que se sitúan los tesauros, los *Topic maps* y las Ontologías.

En las próximas secciones se revisarán algunos de estos conceptos, su utilidad, así como su relevancia actual para las búsquedas en la Web.

1. Las palabras claves en la Web. Las folksonomías

Se denominan folksonomías a los conjuntos de palabras clave incorporadas y asignadas por cualquier internauta para colaborar en la indización de todo tipo de contenidos en un espacio compartido y abierto. La propuesta de este neologismo se atribuye a Thomas van der Wal quién fusionó las palabras *folk* (gente, popular) y taxonomía (Gestión —*taxis*— de la clasificación —*nomos*—), de forma que folksonomía viene a ser etimológicamente "Clasificación (mejor, indización) gestionada popularmente". La asignación de estas etiquetas públicas se realiza sin ánimo de lucro y sin la supervisión de un organismo centralizador, de manera que una de las características de este lenguaje libre es la ausencia de estructuración entre los términos, salvo la formada por el conjunto que describe determinado objeto o concepto, si bien es cierto que cada término tiene sentido de forma individual.

Las folksonomías tienen un gran interés para mejorar la navegación y recuperación de todo tipo de materiales. Ejemplos de folksonomías se pueden ver en las etiquetas para *blogs* en *Technorati, Del.*

Tabla 1. Clasificación de las folksonomías según Hammond (2005)

CREADOR DE LAS ETIQUETAS	PROPIO	AJENO
AJENO	Technorati HTML MetaTags	(Wikipedia)
PROPIO	Flickr	CiteULike Del.icio.us Furl Frassie

CREADOR DEL CONTENIDO

icio.us social bookmarks, para etiquetar sitios Web, o Flickr para fotografías.

Una clasificación popular es la que divide los tipos de folksonomías según la asignación y autor de las palabras-clave y del contenido, como se puede ver en Hammond (2005):

1.1 Beneficios de las folksonomías:

Existen numerosos motivos que explican la popularidad que experimentan estos recursos. Entre los cuales destacan su:

- **Simplicidad en la utilización:** Se trata de una solución simple para usuarios noveles en tareas de indización de contenidos, que no requieren del aprendizaje de un elevado conjunto de reglas para utilizarlas. En los lenguajes facetados, con amplio número de términos y de asociaciones, se complican las decisiones que deben tomarse para indizar un documento, lo que supone un notable coste cognitivo. En la mayoría de los casos esta inversión es muy elevada, por lo que prefieren describir sus documentos con palabras-clave libres (Wal, 2005).
- **Economía:** El carácter social y cooperativo de las folksonomías tiene gran rentabilidad debido a su bajo coste. Los internautas no persiguen lucrarse, sino beneficiarse de mejores búsquedas y navegación, cuantos más usuarios cooperen mayores ventajas se obtienen (Tonkin, 2006).
- **Adecuación al entorno Web:** es el único enfoque posible para indizar los enormes volúmenes de información existentes en la Web (Shirky, 2005), sobre todo cuando la información a indizar no es textual, haciéndose imprescindible la indización manual, es el caso de la indización de videos, fotos, etc.
- **Ejecución de consultas:** las búsquedas pueden ser más específicas, pues los términos asignados por los usuarios tienden a ser concretos. Si bien, uno de los rasgos principales de las taxonomías surge de su cualidad de asociar las verdaderas necesidades de los usuarios con la lengua, no de buscar la precisión (Quintarelli, 2005). Además permite recuperar, como ya se ha comentado,

material multimedia (solucionando en parte el problema de la Internet Invisible).
- **Simplicidad en la gestión:** a diferencia de los lenguajes controlados, los lenguajes libres son más sencillos y económicos por su escaso mantenimiento. Su evolución para incorporar nuevos términos es instantánea al carecer de una autoridad de control, por lo que están siempre actualizadas.
- **Flexibilidad:** la asignación de etiquetas (palabras-clave) a los recursos es flexible, ya que no se trata de un lenguaje precoordinado y no cuenta con un vocabulario definido a priori.

1.2 Desventajas:

Las folksonomías, por el contrario, presentan asimismo notables desventajas, lo cual las relega a ser un instrumento útil en materiales no excesivamente críticos. Los principales inconvenientes se derivan de (Al-Khalifa, 2007; Smith, 2004):

- Ser etiquetas imprecisas, inexactas y ambiguas. Así, la asignación se realiza con criterios subjetivos. Frecuentemente se observan palabras clave que identifican a personas del entorno del usuario, por ejemplo, "Elena" o "Juan".
- Muchas folksonomías solo permiten el uso de unitérminos (p.e. *Del.icios.us*)
- Existen problemas de sinonimia y homonimia que producen imprecisión en las búsquedas debido a un *recall* bajo.

1.3 Otros usos de las folksonomías

Las folksonomías tienen gran interés para estudios sociolingüísticos, pues ayudan a determinar las variaciones que se registran en determinado grupo de hablantes, por lo que permiten realizar estudios de la terminología empleada en un dominio concreto o por una comunidad específica. Una forma de realizar estos estudios consiste en observar la frecuencia de los términos más empleados en determinado contexto y, a su vez, destacar los términos no dominantes (*meta-noise*) que añaden comprensión semántica (Folksonomie, 2004). Además permi-

ten realizar estudios diacrónicos de utilización de determinado término. Un enfoque complementario es analizar los intereses de grupos reducidos de usuarios (Porter, 2004).

Otros estudios interesantes son la denominadas folksontologías, que tratan de crear ontologías a partir de folksonomías (Damme, 2007). Es evidente el valor de muchas de las palabras propuestas a la hora de servir como candidatas a formar parte de algún posible vocabulario controlado, por tratarse de conceptos empleados comúnmente por los usuarios de Internet. Aportan, por tanto, elementos de extracción de palabras libres que, tras su normalización, acabarían conformando los términos de algún tesauro. Para obtener el signo que mejor describe determinado concepto se utiliza la frecuencia de los términos en las folksonomía y las relaciones presentes debidas a la agregación de términos o por relaciones entre sus URL.

No todas las folksonomías son adecuadas para realizar estudios sobre la frecuencia de uso. Así las ontologías genéricas, aquellas en las que muchos usuarios indizan un mismo objeto, permiten un estudio más interesante, un ejemplo de estas ontologías es del.icios.us. Las ontologías específicas (Wal, 2004) permiten solamente a uno o a pocos usuarios clasificar los objetos, este es el caso de Flickr, por lo que su utilidad para medir el consenso de uso de un término es menor. Quizás el mayor inconveniente de esta aproximación es que el consenso tiende a darse más en términos genéricos que específicos, como puede comprobarse empíricamente en la aplicación Image Labeler de *Google*©.

2. Sistematización jerárquica de conocimientos: facetas, categorizaciones y clasificaciones

La necesidad de definir unas categorías generales y jerarquizarlas para así mejorar la organización de la información es una constante a lo largo de los siglos. A continuación se muestra una breve sinopsis que analiza y muestra como ha sido esta evolución.

1) Clasificaciones Clásicas: Según Aristóteles, se pueden definir unas categorías generales en que agrupar todos los objetos, estas categorías son (Aristóteles, 1995): Sustancia, Cantidad, Cualidad, Relación, Lugar, Tiempo, Situación, Posesión, Acción, y Pasión. Posteriormente

Porfirio fue quien propuso disponerlas en forma de árbol, en razón de su género, subtipo y diferencia (Ferrater, 1999). El orden introducido en las categorías aristotélicas por el árbol de Porfirio supuso el primer mapa conceptual, pues representaba gráficamente las relaciones existentes entre los conceptos (Sowa, 2000), (Moreiro, 2006):

Tabla 2. Clasificación de Porfirio

Género Supremo: **Dif. Genérica:**	Sustancia material / inmaterial
Gen. Subalterno: **Dif. Genérica:**	Viviente sensitivo / insensible
Gen. Próximo: **Dif. Específica:**	Animal Racional / irracional
Especie:	Hombre Miguel, Isabel, etc.

En este esquema ya se observa un orden jerárquico. De manera, que el género supremo de los universales es la sustancia material o compuesta, descendiéndose en la escala jerárquica de los universales de acuerdo con el orden marcado en el árbol para las categorías aristotélicas por *Genus* o Género supremo (*Top Term* o Macrodescriptor en un tesauro) y *Species* (Específicos de diferente nivel en un tesauro), hasta llegar a los individuos: Específicos, pasando por Géneros y especies subordinados (descriptores intermediarios o *Middle Terms*) y la Especie espacialísima: Genéricos.

Ramón Llull (Llull, 1998), propuso posteriormente elementos de relación, que se siguen empleando en la actualidad para asociar términos en los tesauros: *Utrum*; *Quid; De quo; Quare; Quomodo; Ubi; Quando; Quantum; Cum quo; Quale,* que además de ser el fundamento lejano de las propuestas lógicas de Port-Royal marcan buena parte de las relaciones asociativas existentes entre los conceptos que un tesauro define (Gayà, 1996). Port-Royal coincidía con Porfirio y Llull en considerar cinco predicables y no cuatro, porque incluían también la especie entre las ideas universales: géneros, especies,

diferencias, propiedades y accidentes. La influencia de estas sistematizaciones ha llegado hasta hoy, pudiéndose afirmar que la sistematización luliana del razonamiento sigue estando vigente en disciplinas como la Inteligencia Artificial (Boden, 1994), las redes semánticas, o la representación del conocimiento (Trillas, 1998).

Kant (Ranjan, 2007) también revisó las categorías Aristotélicas dividiéndolas en: Cantidad (unidad, pluralidad y universalidad); Cualidad (realidad, negación, limitación); Relación (sustancialidad y causalidad); y Modalidad (posibilidad, actualidad y necesidad).

2) Las facetas bibliotecarias: En las bibliotecas se han empleado frecuentemente las facetas para dar mayor flexibilidad a las clasificaciones bibliotecarias. Así la clasificación de Vickery, empleada en bibliotecas inglesas define seis facetas: objeto, parte, propiedad, proceso, operación y agente, y cada faceta se subdivide a su vez mediante una taxonomía. De cualquier modo las facetas que se han empleado con más asiduidad son las definidas por Ranganathan esto es tiempo, espacio, energía, materia y personalidad. Es curioso observar la similitud entre algunas de estas facetas y las categorías discutidas anteriormente.

Según se intentan adaptar las tecnologías a los nuevos formatos la necesidad de definir unas facetas generales permanece, así en el artículo "Clasificaciones Facetadas y Metadatos (I): Conceptos Básicos" de Hassan et al. (2003) se define una clasificación para recursos Web, en el que bitácoras, portales y listas de correo se clasifican según usabilidad, utilización de la Web semántica, temática, tipología e idioma.

En definitiva parece evidente que existen categorías generales que son necesarias para organizar el conocimiento, y que la estructuración jerárquica de estas categorías da un valor añadido a la clasificación original. Esto se ha traspasado a la Web actual en forma de:

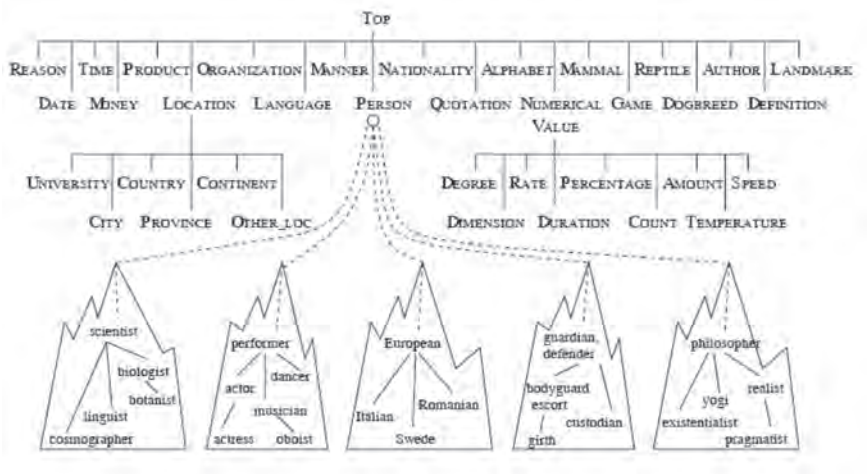
- Vocabularios de Metadatos. Los metadatos son descripciones que facilitan la recuperación, utilización o gestión de recursos de información. Así, los metadatos pueden ser utilizados para organizar recursos electrónicos o facilitar la interoperatividad. Los metadatos suelen estar agrupados en conjuntos de elementos denominados esquemas o vocabularios de metadatos. Cada esquema de metadatos se diseña con una finalidad concreta y se define

dentro de un contexto o espacio de nombres concreto. Dentro de dicho contexto, cada uno de los elementos del esquema posee una definición única. Es habitual el uso de varios espacios de nombres dentro de un mismo esquema, lo que permite el uso de los elementos definidos en otros esquemas, identificados mediante un nombre cualificado (*QName*), esto es: un identificador del esquema, seguido de dos puntos y el nombre del elemento en dicho esquema. Esta característica hace posible la reutilización y difusión de esquemas en la Web. Un ejemplo de estos vocabularios es *Dublin Core*, recurso que establece un conjunto reducido de categorías con las cuales se puede clasificar un recurso, estas son: Elementos (15): Título, Creador, Materia, Descripción, Editor, Colaborador, Fecha, Tipo, Formato, Identificador, Fuente, Lengua, Relación, Cobertura, y Derechos de autor.

- Los directorios de buscadores Web utilizan también unas categorías principales a partir de las cuales palntear las subdivisiones.
- Ontologías de alto nivel o genéricas: un tipo especial de ontología es la denominada de alto nivel. Estas ontologías son útiles para enlazarlas con ontologías de dominio más específicas. De nuevo, un análisis de las mismas muestra fuertes similitudes con las categorías genéricas discutidas en este apartado. Así (Gangemi et al. 2001), las nociones generales proponen agrupar los genéricos en: abstracto, concreto o relación. que la ontología de Cyc, incluida en CycKnowledge Base (Lenat and Guha, 1990), contiene unos 3000 elementos. Estos conceptos han sido agrupados en 43 categorías (*fundamentals, time and dates, spacial relations,* etc.). WordNet por su parte también subdivide su contenido en diferentes clases, entre las raíces están actividad, fauna, artefacto, atributo, cuerpo, conocimiento, comunicación, evento, sentimiento, comida, agrupamiento, lugar, motivo, objeto natural, fenómeno natural, persona, planta, posesión, proceso, cantidad, relación, forma, estado, sustancia y tiempo. De nuevo aparecen algunas constantes como son la sustancia, el tiempo, la forma, la cantidad o el lugar.
- Sistemas de pregunta respuesta: Uno de los desarrollos futuros de la Web consistirá en la evolución de los buscadores desde la

recuperación de documentos a la recuperación de respuestas. Y esto es factible si recuenta con la definición de grandes categorías de consultas. Los sistemas pregunta-respuesta también han podido categorizar a estas alcanzando un conjunto de categorías análogas. Un ejemplo se puede ver en la taxonomía de Pasca (2003)

Figura 1: Taxonomía de Pasca



Los sistemas OLAP (*On-Line Analytical Processing*), tienen también definidas estas categorías. Se trata de almacenes de datos estructurados para facilitar el procesamiento analítico de apoyo a la toma de decisiones estratégicas. De esta manera, los datos se organizan en torno a hechos (por ejemplo, ventas), cada hecho tiene unos atributos/medidas (por ejemplo: importe, cantidad, número de clientes, etc.) con mayor o menor granularidad a la hora de reflejar sus dimensiones (por ejemplo; CUÁNTO se ganó con la venta, CUÁNDO fue venta, DÓNDE se realizó venta, QUÉ se vendió, etc.) (Hernández, et al., 2004). Dicho de otro modo, de nuevo se detectan unos presupuestos universales que son constantes en todos los esquemas: como las facetas: cuándo, qué, dónde, etc, o los atributos, que suelen responder al cuánto.

2.1 Taxonomías

El concepto de taxonomía proviene del siglo XIX cuando con ellas se ordenaban, describían y clasificaban los seres vivos para las Ciencias naturales, partiendo de la especie como unidad de clasificación. Algo parecido sucedía con la Terminología científica pues, como consecuencia de la internacionalización progresiva de la ciencia, los científicos se esforzaban en sus reuniones por ordenar las nomenclaturas dentro de estructuras taxonómicas.

En la actualidad, las taxonomías se aplican con otro sentido en el mundo empresarial e institucional para organizar y gestionar los recursos de información digitales que como organizaciones complejas alojan en sus servidores Web, buscando categorizarlos, hojearlos y navegar por ellos. Así, se emplean términos autorizados en cada institución, con definiciones que usa una organización para clasificar sus contenidos (Corcoran, 2002). Los usuarios clasifican las materias dentro de jerarquías para hacer fácil la búsqueda de los recursos de información (Zhongong 2007).

Una taxonomía organiza no sólo los contenidos propios de una organización, sino también los servicios que ofrece, sus productos y cuanto se deriva de la experiencia y datos sobre los recursos humanos. De esta manera, resulta una red semántica de conceptos interrelacionados para cubrir con validez específica las necesidades empresariales y la forma con que los trabajadores se relacionan con la información (Conway, 2002). La aplicación práctica de las taxonomías se consigue cuando las usamos para navegar en la Web. Permiten a los usuarios acceder a ítems de interés específico enlazando recursos a partir de sus correspondientes categorías que posibilitan ir estrechando sus campos de búsqueda. Un ejemplo son los directorios tipo *Yahoo©* o *DMOZ*.

Las taxonomías funcionan dentro de un contexto específico, se basan en razones internas. Son flexibles y fáciles de modificar por funcionar tanto por jerarquía como por facetas. Integran nuevas áreas de interés y modifican fácilmente su estructura de acuerdo con las necesidades de cada momento. Como característica de los términos que las componen, se resalta que son categorías representadas en entradas

etiquetadas y orientadas al usuario. Por todo ello, podemos afirmar que las taxonomías generan sus estructuras jerárquicas de acuerdo con un contexto y unos usuarios determinados.

2.2. Las taxonomías en los lenguajes documentales: tesauros y ontologías

En 1876, Dewey (Dewey, 1979) presentó su *Clasificación decimal* que marcó el camino a seguir por los sistemas clasificatorios, precoordinados y de estructura jerárquica, y las *Rules for a dictionary catalog* de Charles Cutter (Cutter, 1962) que se anticipó a las listas de encabezamientos de materia y, de alguna forma, a los lenguajes controlados. Las teorías de Cutter han tenido mayor proyección hacia los lenguajes controlados que las clasificaciones debido a su carácter precoordinado y a su estructura asociativa, así como al control de vocabulario de aplicación específica a los conceptos y a la facilidad de uso para el usuario, frente a la rigidez arbórea de los sistemas clasificatorios, surgidos de la idea de Dewey para analizar el conocimiento humano. Más adelante apareció la Clasificación Decimal Universal (CDU), adaptación de la ideada por Dewey, para relacionar los conceptos del *Repertorio universal* de La Fontaine y Otlet por jerarquía, similitud o diferencia, por lo que, además de mostrar características de taxonomía, tenía en cuenta las asociación.

Fue durante la década de 1960 cuando, ante la aparición de las bases de datos como una de las aplicaciones para enfrentar el creciente número de publicaciones científicas y técnicas, el tesauro pasó a ocupar un lugar protagonista en la recuperación de los documentos. Se requerían soluciones que los sistemas tradicionales eran incapaces de suministrar, pues ahora había que representar los conceptos contenidos en los documentos, así como las relaciones existentes entre los conceptos, en una forma de lenguaje estandarizada, obtenida mediante el control de sinónimos y con una estructura sintáctica más simplificada que la del lenguaje natural, que representara el contenido informativo de los documentos de forma normalizada. Cualquier tesauro parte de una categorización del dominio que cubre y, por lo tanto, de una taxonomía del conocimiento temático que, en este sen-

tido, es una terminología jerarquizada. Las taxonomías están presentes en todos los *Esquemas, Tesauros, Modelos conceptuales y Ontologías*. Entendiendo por Taxonomía la clasificación o categorización de un conjunto de términos (en nuestro caso descriptores) de forma jerárquica, que establece una relación esquemática entre los objetos de generalización-especialización (Daconta, 2003).

2.3 Web Semántica y Ontologías

En 1998, Tim Berners-Lee publicó en la página principal del *Worl Wide Web Consortium* (W3C en adelante) su conocido artículo *Roadmap to the Semantic Web.* En dicho artículo se introducía por primera vez el término Web Semántica (*Semantic Web*) destacándose la necesidad de expresar la información de forma que ésta fuera procesable por máquinas. El artículo presentaba un conjunto de pasos hasta lograr una Web en la que el razonamiento fuera automático, llevado a cabo por máquinas, y distribuido. En 2001, Berners-Lee entregó un nuevo artículo *The Semantic Web* en el que presentó las principales características que tendrá la futura Web convencido de que las máquinas facilitarán nuevas prestaciones al mejorar su capacidad de procesar y comprender la información dispersa por la Web. Como solución a las limitaciones semánticas de la actual Web, propuso hacer procesable de forma automática el contenido de la Web, llegando así a la definición de Web Semántica:

The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. De modo que la Web Semántica no sería una nueva Web sino la extensión de la existente, mediante la adición de metadatos que describan la semántica de las páginas de forma procesable por máquinas.

Por ello, en la Web Semántica el usuario final de las ontologías no son las personas sino las aplicaciones informáticas. La forma de hacer que cualquier aplicación pueda leer y procesar los datos de un repositorio estableciendo identificadores de los documentos basados en URLs y un lenguaje con sintaxis estandarizada denominado XML (W3C, 2006). Además es necesario expresar el conocimiento de ma-

nera uniforme, algo que se consigue mediante tripletas expresadas en RDF (W3C, 2005), por último, se necesitan vocabularios controlados que aporten semántica y que denominen los conceptos de forma no ambigua, permitiendo la navegación a través de sus estructuras. Esto es posible gracias a las ontologías y a los vocabularios de metadatos. La semántica basada en vocabularios controlados y ontologías permitiría a las aplicaciones interoperar con cualquier repositorio de datos, convirtiendo la Web en una gran base de datos.

2.3.1 Ontologías

Aunque con cierta confusión en su concepto actual, una ontología o conjunto de conocimientos representado formalmente se fundamenta en la conceptualización de los objetos y entidades que existen en un área de interés, así como en las relaciones que existen entre ellos (Genesereth, 1987). Una ontología sirve, pues, para categorizar los conceptos propios de un dominio y sus relaciones (Gruber, 1994), estableciéndose, por tanto, como especificación explícita de una (o una parte de una) conceptualización que incluye un vocabulario organizado de términos.

Tanto tesauros como ontologías, tienen un conjunto de etapas comunes en su construcción entre los que destacan (Sanchez-Cuadrado, 2007):

1) Identificación de conceptos relevantes del dominio
2) Selección de los términos que mejor representan los conceptos
3) Creación de una taxonomía de conceptos
4) Identificación de otras relaciones no jerárquicas

Una ontología puede tener además una mayor riqueza semántica, ya que puede contener funciones calculadas, constantes, propiedades, instancias, axiomas o restricciones. La necesidad de esta mayor riqueza semántica se justifica en la capacidad de algunas ontologías de realizar inferencias o razonamientos. Así pues, la construcción de ontologías pasa por un proceso de jerarquización, que se efectúa mediante la subdivisión o agrupación de clases hasta alcanzar una taxonomía bien organizada (Noy, 2000). De forma que la organización taxonó-

mica aparece como condición mínima en la superestructuración de los lenguajes documentales de carácter combinatorio, formando la agrupación primera de términos en los tesauros y en las ontologías, y alcanzando niveles de esencialidad en los de carácter clasificatorio.

2.3.2 Implantación Actual de la Web Semántica

Lamentablemente, la Web Semántica, operativa desde 1999, no ha tenido el éxito esperado. Una búsqueda en *Google*© muestra que, en junio de 2007, existían 2,770.000 documentos con extensión RDF, 41,500 con OWL, 2,330 XTM, 4,540,000 RSS y 212,000 ATOM, es decir 7,5 millones de documentos en un conjunto de 10,000 millones de documentos. Con Swoogle [1], un recuperador especializado en la Web Semántica, los resultados no son mucho mejores: 158,000 documentos semánticos conteniendo el término RDF y 2,323,857con el término OWL.

Las causas son diversas, pero debemos considerar entre las principales:

- Falta de legibilidad de los lenguajes RDF y OWL, lo cual supone un cuello de botella para que los expertos validen las ontologías (Gómez-Pérez, 2004). En 2005, Mika (Mika, 2005) subrayo la importancia que tienen los usuarios (denominada Dimensión Social en el artículo original) para la aceptación de la Web Semántica. De hecho la Web Semántica tiene diferentes grados de complejidad en la creación de recursos, y esta complejidad es inversamente proporcional a la proximidad al usuario (Fig.2). En la figura 2 se muestra gráficamente este hecho, así existe una tendencia que provoca que cuando se incrementa la complejidad en la representación semántica se produce una disminución en la dimensión de contacto con el usuario. Esta dimensión social engloba diferentes afectos como usabilidad, legibilidad o necesidad de formación previa para su interpretación.

1 Swoogle http://swoogle.umbc.edu/

- Escasez de herramientas que faciliten la creación de documentos semánticos mediante formularios usables, un ejemplo de un entorno más amigable se puede ver con Protégé o con Tabulator. [2]
- La migración de folksonomías a folkontologías es un tema aún por desarrollar, aunque ya existen estudios (Damme, 2007; Matsuo, 2006). Básicamente, esta migración se hace mediante herramientas estadísticas (Bagelman, 2006) y lingüísticas, o incluso proponiendo una normalización para asignar etiquetas (Xu, 2006).
- La incorporación de técnicas semiautomáticas para la creación de Sistemas de Organización del Conocimiento basadas en PLN y Minería de Datos (Sánchez-Cuadrado, 2007), ya que la carencia de estos recursos, junto con la lentitud en su creación dificulta la implantación de la Web Semántica.
- Presencia de duplicidades en los Vocabularios de Metadatos y Ontologías, lo cual provoca la desconfianza y confusión del usuario que no sabe cuál es el vocabulario idóneo o más generalizado. Como ejemplo están los vocabularios de metadatos para expresar tesauros. Actualmente existen, entre otros, el SKOS-Core[3] del W3C, los PSI[4] de los Topic Maps, Zthes[5] y MADS[6].

---

2 Tabulator: Async Javascript and Semantic Web. http://dig.csail.mit.edu/2005/ajar/release/tabulator/0.8/tab.html

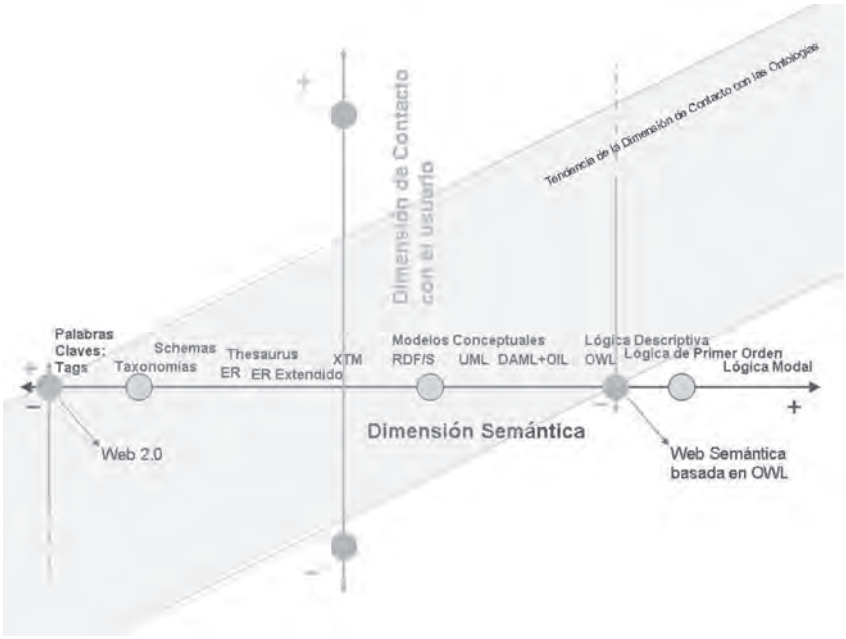
3 SKOS Core Guide http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20050510/

4 Published Subject Indicators For Modelling Thesaurii http://www.techquila.com/psi/thesaurus/

5 Zthes - specifications for thesaurus representation, access and navigation http://zthes.z3950.org/

6 Metadata Authority Description Schema (MADS) http://www.loc.gov/standards/mads/

Figura 2. Relación entre la Complejidad de la Web Semántica y la proximidad al usuario



3. Una visión integradora de los vocabularios para la Web

Como consecuencia del recorrido hecho en este texto, aparece con claridad que:

- Las folksonomías tienen su propio espacio en la Web, y que si bien sería deseable un control del vocabulario para evitar ambigüedades semánticas e inexactitudes en la asignación de términos, su reemplazo parece complicado debido a la simplicidad de uso que presentan.
- Los vocabularios controlados y su encaje en la Web Semántica necesitan un gran desarrollo, pero la falta de apoyo popular hace que este se ralentice.

Como consecuencia, parece evidente que la única forma de popularizar la Web Semántica y de evitar los problemas derivados de la ambigüedad inherente a las folksonomías sería implicando a los usuarios en el desarrollo y empleo de los vocabularios controlados, pero de una manera simplificada. La mayoría de los esfuerzos en esta línea (Fumero, 2007) tratan de mejorar los interfaces y la usabilidad, y de mostrar a los usuarios los beneficios de estos vocabularios. A continuación se muestran ocho elementos en los que habría que avanzar para conseguir este propósito:

3.1 Analizar cuales son en la actualidad los documentos de la Web Semántica Social

Posiblemente, en la actualidad, el enfoque más próximo a la Web Semántica Social son los documentos RSS.[7] Estos documentos, que se cuentan por millones en Internet, aúnan la sindicación de contenidos (un objetivo típico de la Web 2.0) con la expresión de dicha sindicación en RDF (en el caso de RSS 1.0) o XML (en el caso de RSS 2.0). Sin duda, como en el caso del vocabulario Dublín Core (DCMI, 2007), el éxito se ha basado en dos factores: la aparición de lectores usables dirigidos al usuario final y la simplicidad del vocabulario de metadatos subyacente.

Las etiquetas Meta de HTML y los Microformatos[8] son dos soluciones que tienen respaldo popular, si bien son poco ambiciosos para crear la Web Semántica, muestran que la solución para implantar esta Web debe ser simple. La simplificación (Senso, 2007) se puede dar de diferentes formas, por ejemplo, prescindiendo de los lenguajes formalizados para ontologías a favor de expresar el contenido mediante documentos XHTML (una solución denominada microformatos), esta opción dado su carácter de solución local, no consigue uno de los principales objetivos de la Web Semántica, la interoperabilidad. Por el contrario, la interoperabilidad no se ve mermada en determinados documentos semánticos que han tenido un gran éxito gracias a la sim-

7 Wikipedia: RSS http://es.wikipedia.org/wiki/RSS

8 Wikipedia: Microformatos http://es.wikipedia.org/wiki/Microformatos

plicidad del vocabulario subyacente, un ejemplo son la sindicación de contenidos mediante RSS.

3.2 Promover la creación de ontologías de dominio por usuarios expertos

La creación de ontologías de forma colaborativa es un tema complejo, ya que las ontologías se basan en el consenso para determinar los conceptos e interrelaciones relevantes entre los mismos, y este consenso se dificulta a medida que aumenta el número de usuarios. Por otra parte, cuanto más específico sea el dominio más complicado es identificar expertos dispuestos a colaborar.

Una posible solución es que los organismos públicos incentiven y financien la creación de estos recursos. La opción de que exista capital privado para crearlos no parece realista, ya que el resultado deberá ser un producto gratuito y reutilizable por terceros.

3.3 Establecer mecanismos para centralizar los documentos semánticos en un repositorio común y eliminar duplicidades innecesarias

Actualmente existen al menos cuatro vocabularios de metadatos para expresar un KOS tipo tesauro. El SKOS-Core[9] del W3C, los PSI[10] de los Topic Maps, Zthes[11] y MADS[12]. La duplicidad para crear de forma descontrolada vocabularios controlados provoca la desconfianza y confusión del usuario que no sabe cuál es el vocabulario idóneo o más generalizado.

---

9 SKOS Core Guide http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20050510/

10 Published Subject Indicators For Modelling Thesaurii http://www.techquila.com/psi/thesaurus/

11 Zthes - specifications for thesaurus representation, access and navigation http://zthes.z3950.org/

12 Metadata Authority Description Schema (MADS) http://www.loc.gov/standards/mads/

3.4 Mejorar la dimensión de contacto en la utilización de documentos RDF-OWL

La dificultad de uso de algunos lenguajes como OWL no anima a los usuarios a colaborar. De hecho, según Gómez-Pérez et al. (2004) la validación de la ontología OWL, por parte de los expertos en el dominio, es uno de los cuellos de botella en la creación de ontologías.

El problema ha sido atajado en anteriores ocasiones con algunas propuestas como es el caso de la Wikipedia en el que las mejoras se realizan mediante diferentes perfiles para crear, editar o validar los nuevos contenidos, pero bajo un interface amigable, o como pueda ser la creación de Wikis de noticias mediante tags que sean codificados en RDF de forma transparente (Fuentes, 2007).

Algunas aplicaciones de Altova (SemanticWorks[13]) y Microsoft (InfoPath[14]) o los ya mencionados formularios para edición de RSS facilitan la incorporación de instancias en RDF

Otro problema relativo a este punto es la mejora de los interfaces para navegar por estos documentos, algo que intenta realizar el proyecto Tabulator.[15]

3.5 Desarrollar herramientas para convertir las folksonomías en KOS más complejos

Algunos recursos Web están siendo motivo de análisis para su evolución hacía KOS complejos, un ejemplo son las folkontology (Damme, 2007), que estudian mecanismos de evolución a partir de una folksonomía.

---

13 Altova SemanticWorks - visual RDF and OWL editor that autogenerates RDF/XML and N-Triples http://www.altova.com/products/semanticworks/semantic_web_rdf_owl_editor.html

14 Informacion de Producto de Infopath 2003 http://www.microsoft.com/spain/office/products/infopath/default.mspx

15 Tabulator: Async Javascript and Semantic Web http://dig.csail.mit.edu/2005/ajar/release/tabulator/0.8/tab.html

Otro ejemplo, sobre esta evolución se da en la Wikipedia[16],[17]. Sin duda, la incorporación de las nubes de tags con herramientas lingüísticas y estadísticas puede ayudar en este proceso. Un ejemplo es Piggy Bank,[18] una aplicación que captura localmente las etiquetas de los RDF visitados, para organizarlo en una ontología local, y que puede ser compartida en el Semantic Bank (Senso, 2007). La aplicación permite poner *tags* a las URL seleccionadas para recuperarlas posteriormente. Un proyecto muy similar ha sido propuesto por el W3C bajo la denominación de Annotea,[19] instalable mediante un plugin en Firefox denominado Annozilla. El problema de ambas aproximaciones son las inherentes a las *tags*, es decir, la sinonimia y la polisemia.

3.6 Promover una arquitectura de colaboración con la implicación de los usuarios

El usuario debe tener claro que su colaboración para mejorar el recurso no implicará la obligación de pagar un sobreprecio para su futuro uso del recurso mejorado. Además, el usuario debe ser consciente de los beneficios que la Web Semántica le supondrá a corto plazo.

3.7 Establecer mecanismos para mejorar la fiabilidad de las búsquedas

El spam semántico será una realidad cuando se comience la transición hacia la Web Semántica. Un gran número de motores Web no utiliza en el posicionamiento las etiquetas Meta de HTML porque inicialmente las técnicas fraudulentas de optimización las utilizaron con asiduidad. Así, parece evidente que se deberán establecer sitios de confianza para evitar la desconfianza del usuario.

---

16 Harvesting Wiki Consensus - Using Wikipedia Entries citeseer.ist.psu.edu/747700.html

17 Ontoworld.org http://ontoworld.org/wiki/Main_Page

18 SIMILE- My Piggy Bank http://simile.mit.edu/wiki/Piggy_Bank

19 W3C: Annotea Project http://annotest.w3.org/

3.8 Reducir la desconfianza de las empresas

Si los repositorios de todas las empresas de un dominio exportan los datos bajo el mismo modelo y vocabulario de metadatos, se incrementará la interoperabilidad para consultar distintos repositorios, lo cual facilitaría que la comparación de precios y servicios se pueda automatizar en diferentes grados. Esto puede provocar la desconfianza inicial de determinadas empresas, aunque es previsible que esta tendencia se reduzca cuando se alcance determinado peso crítico de la Web Semántica, ya que se correrá el riesgo de quedar relegado fuera del mercado.

Conclusiones

El enorme número de documentos que pueblan actualmente Internet hace necesario el empleo de diferentes lenguajes que varían en su grado de estructuración. Cada uno de estos lenguajes tiene su sitio dependiendo de la funcionalidad perseguida y de los recursos disponibles. Así, para describir los materiales multimedia no críticos desde el punto de vista de su valor se han impuesto las folksonomías. Las taxonomías y las facetas han ocupado su lugar para organizar portales Web. Las ontologías se hacen necesarias para hacer de la Web una gran base de datos en la que cualquier aplicación pueda comprender la información allí disponible.

Por otro lado, los lenguajes han tenido que ser transformados, creándose relaciones adaptables a cada dominio, así como la ampliación del concepto de tesauro de descriptores mediante la admisión de nuevas categorías gramaticales que han enriquecido con nuevos matices la semántica del mapa conceptual y, desde luego, aumentando con nuevas categorías las relaciones interconceptuales que han alcanzado incluso a los recursos de información, y que han extendido las posibilidades de asociación conceptual aproximándolas a la riqueza casuística del lenguaje natural. Sin embargo, la razón fundamental de las relaciones entre los términos de los lenguajes documentales sigue basándose en la estructuración jerárquica, tal como se establece para la terminología propia de un campo científico. La clasificación ha sido

estudiada por los principales autores filosóficos cuando se han acercado a poner las bases retóricas de los discursos. Esos principios siguen siendo fundamentales a la hora de organizar los lenguajes combinatorios, e incluso en sus evoluciones, pues si la diferencia entre los tipos de lenguaje parte de las posibilidades aumentadas de asociar términos, e incluso del concepto de término que se tenga, lo común sigue siendo la organización jerárquica en taxonomías, que afecta comúnmente a las clasificaciones jerárquicas, a los lenguajes combinatorios y a las propias ontologías.

Otra alternativa es la que ha aparecido con la asignación de palabras libres por los propios usuarios a los contenidos digitales que la red difunde y cuyo contenido sería imposible de analizar de otro modo. Las Folksonomías vienen a cubrir las necesidades de indización de los documentos Web que no son atendidos por los grandes servicios de pago o públicos. En este sentido suponen una solución popular al problema de los legítimos intereses de grupo en los documentos situados fuera de los cauces de circulación controlada o económicamente muy productivas.

Las folksonomías han venido a renovar las formas de indizar, pues han distribuido su responsabilidad entre los usuarios y han impuesto métodos descentralizados, alejados de cualquier jerarquía sistemática. Si bien actualmente se hacen necesarias técnicas que aproximen las folksonomías, propias de la Web 2.0, a unos lenguajes controlados, eliminándose problemas propios de los lenguajes libres, como la sinonimia, la homonimia y la ausencia de niveles de estructuración de los términos entre si. En el documento se exponen algunas recomendaciones para conservar la percepción de simplicidad de las palabras claves pero dentro de un entorno de gestión orientado a la implantación de lenguajes controlados.

Agradecimientos

Los resultados del presente artículo han podido obtenerse gracias a la ayuda financiera de la Comunidad de Madrid a través de su programa 2007/04055/001

BIBLIOGRAFÍA

(Al-Khalifa, 2007) Al-Khalifa, H.- Automatic document level semantic metadata annotation using folksonomies and domain ontologies. 2007. http://eprints.ecs.soton.ac.uk/14181/01/Hend_Thesis.pdf

(Aristóteles, 1995) Aristóteles.- *Tratados de Lógica. Organon.* Madrid: Gredos, 1995. v. 1: 45.

(Begelman, 2006) Begelman, G.; Keller, P. and Smadja, F.- Automated Tag Clustering: Improving search and exploration in the tag space. WWW2006, May 22–26, 2006, Edinburgh, UK. http://www.rawsugar.com/www2006/20.pdf

(Berners-Lee, 1999) Berners-Lee, T. A roadmap to the Semantic Web. Disponible en: <http://www.w3.org/DesignIssues/Semantic.html>. [Consultado en junio de 2007].

(Berners-Lee, 2001) Berners-Lee, T.;Hendler, J.; Lassila, O.- The Semantic Web. Scientific America, 2005, vol. 284, num. 5, pp. 34-43. ISSN 0036-8733.

(Boden, 1994) Boden, M. (comp.).- *Filosofía de la Inteligencia Artificial*. México: Fondo de Cultura Económica, 1994.

(CiteULike) CiteULike.com. A free online service to organise your academic papers, en www.citeulike.org/. [consulta 18-03-2007].

(Conway, 2002) Conway, S. y Sligar C.- Building taxonomies, en su *Unlocking knowledge assets*. Redmont: Microsoft Press, 2002: 105-124.

(Corcoran, 2002) Corcoran, M.- Industry insights: taxonomies, hope or hype?, en *Online*, 2002, 26, nº 5: 76-79.

(Cutter, 1962) Cutter, Ch.- *Rules for a dictionary catalog*. 4th ed. London: Chaucer House, 1962.

(Daconta, 2003) Daconta, M.; Obrst, L; y Smith, K.- *The Semantic Web. A guide to the future of XML, Web services, and Knowledge management*. Indianapolis: Wiley Publishing, 2003: 145.

(Damme, 2007) Damme, C. et al.- Ontology: An Integrated Approach for Turning Folksonomies into Ontologies. 2007. http://www.heppnetz.de/files/vandammeheppsiorpaes-folksontology-semnet2007-crc.pdf

(DCMI, 2007) DCMI. Dublin Core Metadata Initiative. 2007 http://es.dublincore.org/

(Dewey, 1979) Dewey, M.- *Decimal classification and relative index*. 19th ed. Albany: Forest Press, 1979.

(Ferrater,1999) Ferrater Mora, J.- *Diccionario de filosofía*. Nueva ed. rev., aum y act. por Josep-María Terricabras; supervisión de Priscilla Cohn Ferrater. Barcelona: Ariel, 1999. v. 1: 49.

(Folksonomie, 2004) Folksonomie. Many 2 many. A group weblog on social software, en www.corante.com/many/archives/2004/08/25/folksonomy.php. 2004 [Consulta 15-03-2007].

(Fuentes, 2007) Fuentes, D. et al.- CoolWikNews:More than meets the eye in the XXI century journalism. Emerging technologies form semantic work environments: techniques, methods, and applications. Idea group, Germany, 2007

(Fumero, 2007) Fumero, A Roca G y Encinar, J.- Web 2.0. Colección Fundación Orange. 2007. http://www.fundacionauna.com/areas/25_publicaciones/indice_web2.asp

(Gayà, 1996) Gayà, Jordi. El arranque filosófico del Ars luliana. Constantes y fragmentos del pensamiento luliano, en Domínguez, F. y Salas, J. de (eds).- *Actas del simposio so-*

bre Ramon Llull en Trujillo (1994). Tübingen: Max Niemeyer, 1996: 1-8.

(Genesereth, 1987) Genesereth & Nilsson.- *Logical Foundation of Artificial Intelligence.* Los Altos (Ca.): Morgan Kaufmann Publishers, 1987.

(Gómez, 2004) Gómez, F.- Grounding the ontology on the semantic interpretation algorithm, en *Proceedings of the Second International WordNet Conference.* Masaryk University, Brno, 2004: 124-129.

(Gómez-Pérez, 2004) Gómez-Pérez, A. et al.- Ontological Engineering. Springer, London, 2004.

(Gruber, 1994) Gruber, T.R.- Toward principles for the design of ontologies used for knowledge sharing, en Guarino, N. y Poli, R. (Eds.).- *International Workshop on Formal Ontology, Padova, Italy. Revised August 1993.* Publicado en Guarino, N. y Poli, R. (Eds.).- *International Journal of Human-Computer Studies.* Special issue on Formal Ontology in Conceptual Analysis and Knowledge Representation 1994. Disponible en http://ksl-eb.stanford.edu/KSL_Abstracts/KSL-93-04.html [consulta 14-03-2003].

(Hammond, 2005) Hammond, T., T. Hannay, B. Lund and J. Scott.- Social Bookmarking Tools (I): A General Review. D-Lib Magazine, 2005, 11(4).

(Hassan, 2003) Hassan Montero, Y.; Martín Fernández, F.J. y Martín Rodríguez, O.- Clasificaciones Facetadas y Metadatos (I): Conceptos Básicos. No Solo Usabilidad [en línea], 2. (28 Febrero 2003). 2003. http://www.nosolousabilidad.com/articulos/clas_facetadas1.htm

(Hernández, 2004) Hernández Orallo, José.; Ramirez Quintana, Mª José; Ferri Ramirez, César - Introducción a la Minería de Datos. Madrid. Pearson Prentice Hall, 2004.

(Matsuo, 2006) Matsuo, Y. et al.- Spinning multiple social networks for semantic web. In Proc. AAAI-06, 2006.

(Mika, 2005) Mika, P.- Ontologies are us: A unified model of social networks and semantics. Proceedings of the 4th International Semantic Web Conference (ISWC 2005), LNCS, 3729, Springer-Verlag, 2005. http://www.cs.vu.nl/~pmika/research/papers/ISWC-folksonomy.pdf

(Moreiro, 2006) Moreiro González, José Antonio; Morato Lara, Jorge; Sánchez Cuadrado, Sonia; Rodríguez Barquín, Beatriz A.- Categorización de los conceptos en el análisis de contenido: su señalamiento desde la Retórica clásica hasta los *Topic Maps,* en *Investigación Bibliotecológica: archivonomía, bibliotecología e información,* 2006, 20, nº 40: 13-31.

(Noy, 2000) Noy, N. y McGuinnes, D. L.- *Ontology development 101: a guide to creating Your First Ontology* [Página web]. Disponible en: http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html [Consulta 14-03-2007].

(Porter, 2004) Porter, J.- Controlled vocabularies cut off the long tail, en http://bokardo.com/archives/controlled_vocabularies_long_tail/. 2004. [consulta 18-03-2007].

(Ranjan, 2007) Ranjan Hatua, Sudip. Categories in Classification. www.geocities.com/sudiphatua/class.html

(Sánchez-Cuadrado, 2007) Sánchez-Cuadrado, S.- Definición de una metodología para la construcción automatizada de un sistema de organización del Conocimiento. 2007. Tesis Doctoral Universidad Carlos III.

(Senso, 2007) Senso José A.- Navegadores semánticos o semantizar el navegador. ThinkEPI. 2007. http://www.thinkepi.net/repositorio/navegadores-semanticos-o-semantizar-el-navegador/

(Shirky, 2005) Shirky, C.- Folksonomies + controlled vocabularies, en http://www.corante.com/many/archives/2005/01/07/folksonomies_controlled_vocabularies.php. 2005, [Consulta 14-03-2007].

(Smith, 2004) Smith, G.- Folksonomy: social classification, http://atomiq.org/archives/2004/08/folksonomy_social_classification.html. [Consulta 14-03-2007].

(Sowa, 2000) Sowa, J. F.- *Knowledge representation: Logical, Philosophical and Computational Foundations*. Pacific Grove: Brooks / Cole Thompson Learning, 2000.

(Tonkin, 2006) Tonkin, Emma.- Folksonomies: the fall and rise of plain-text tagging, en *Ariadne*, 47. http://www.ariadne.ac.uk/issue47/tonkin/intro.html [Consulta 07-07-2007].

(Trillas, 1998) Trillas, E.- *La inteligencia artificial*. Madrid: Debate, 1998.

(W3C, 2005) W3C- Primer: Getting into RDF & Semantic Web using N3. 2005. http://www.w3.org/2000/10/swap/Primer

(W3C, 2006) W3C- Extensible Markup Language (XML). 2006 http://www.w3.org/XML/

(Wal, 2004) Wal, Thomas Van der.- Explaining and Showing Broad and Narrow Folksonomies: http://www.vanderwal.net/random/entrysel.php?blog=1635. 2004. [consulta 23-07-2005].

(Wal, 2005) Wal, Thomas Van der.- *Off the Top: Folksonomy Entries*, en http://www.vanderwal.net/random/category.php?cat=153, 2 de noviembre de 2005 (Consultado el 11 de noviembre de 2005)

(Xu, 2006) Xu, Z. et al.- Towards the Semantic Web: Collaborative Tag Suggestions. WWW2006, May 22–26,

2006, Edinburgh, UK. http://www.rawsugar.com/www2006/13.pdf

(Z 39.19-2005) Z 39.19-2005: NISO (National Information Standard Organization.- *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies.* NISO Press, Bethesda, Maryland, U.S.A., 2005: http://www.niso.org/standards/resources/Z39-19-2005.pdf

(Zeng, 2004) Zeng Lei, Marcia, y Chan, L. M.- Trends and issues in establishing interoperability among knowledge organization systems, en Journal of the American Society for Information Science and Technology, 2004, 55, nº 5: 377-395.

(Zhongong, 2006) Zhongong, W., Chaudry, A. S., y Khoo, C.- Potential and prospects of taxonomies for content organization, en *Knowledge Organization,* 2006, 33, nº 3: 160-169.