

La representación de los contenidos digitales: de los tesauros automáticos a las folksonomías

J.A. Moreiro González [jamore@bib.uc3m.es]
Universidad Carlos III de Madrid

Resumen: Reflexión sobre las dificultades que aparecen cuando se quieren explotar las potencialidades de los contenidos digitales y la respuesta dada por los lenguajes de representación documental. Desde la adaptación a los requerimientos de la información positivista se recorre el camino evolutivo de los lenguajes documentales hasta alcanzar las tendencias actuales condicionadas por la pretendida Web semántica. Se recorren las tendencias en la representación mediante lenguajes documentales al servicio de un concepto postmoderno de información, por un recorrido que parte de los tesauros como referente hasta alcanzar las ontologías como aspiración, pasando por los tesauros automáticos, los tesauros de verbos y los Topic Maps, sin olvidarse de la vigencia de los sistemas basados en palabras-clave en la Web, las folksonomías.

Abstract: Reflection on the difficulties that appear when wants to exploit the potentialities of the digital contents and the answer given by languages of documental representation. From the adaptation to the requirements of the positivist information it looks through the evolutionary way of documentary languages to reach current trends conditioned by claimed semantic web. It looks over trends in representation through documentary languages in the service of a postmodern concept of information, surveying from thesauri as reference to reach ontologies as aim, crossing automatic thesauri, verb's thesauri and Topic Maps, not forgetting life of systems based on keywords in Web, folksonomies.

Palabras-clave: Lenguajes documentales; taxonomías; ontologías; tesauros; Topic maps; tesauros de verbos; folksonomías; Web semántica.

Keywords: Documentary languages; taxonomies; ontologies; thesauri; Topic maps; Verbs Thesauri; folksonomies; Semantic Web.

1. Introducción

Hacia mediados del siglo XX, con la creciente especialización del conocimiento, empezó a ganar fuerza la idea del concepto como núcleo organizador de los lenguajes documentales. No se trataba ya de dar un orden físico a los impresos, si no de guiar hacia los conceptos o los significados de la información. Cuando comenzaba la década de los sesenta se impuso la idea de descriptor como forma de alta significación que representa a un concepto o idea (Mooers, 1963). En cuanto unidad constitutiva de un tesoro, el descriptor, es decir, el concepto y sus relaciones, se determinaron como el elemento clave para elaborar lenguajes documentales. El tesoro se convirtió en el lenguaje modelo para posibilitar la recuperación de los documentos en los sistemas de información. Los soportes actuales no han podido decidir cual de los lenguajes documentales es el más idóneo ya 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:

- 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).

- 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.

Sin embargo, la mayoría de los sistemas de representación se generó antes de que existiera el ciberespacio. Y en la actualidad, la inteligencia artificial y la Web conocida como “semántica”, que apareció hace más de una década y que está dirigida por un consorcio de grandes compañías (Yahoo, Google, AOL, IBM, Microsoft, etc.) no logran los avances esperados, limitándose a la automatización de las operaciones lógicas para sacar el mejor partido de los ordenadores, pero no han alcanzado la nueva matriz simbólica que debería de corresponderse con el calificativo de semántica (Lévy, 2007). Así pues, pese al refinamiento técnico, los documentalistas del siglo XXI se enfrentan al problema de inventar, adaptar y mejorar una nueva generación de sistemas representativos que se ajuste a la potencia de tratamiento ya disponible.

Aparecen serias dificultades a la hora de explotar las potencialidades de lo digital. El primer conjunto de dificultades se origina en la pluralidad y subdivisión de los sistemas representativos, como consecuencia de la gran variedad de lenguas naturales: con recíproca incompatibilidad e inadaptación a los numerosos sistemas de identificación, heredados de la era bibliográfica (no fueron concebidos para utilizar la interconexión y la potencia de cálculo del ciberespacio). En paralelo contamos con una enorme pluralidad e incompatibilidad de taxonomías, ontologías, terminologías, tesauros, y sistemas de clasificación (vinculados a diferentes culturas, tradiciones, teorías y disciplinas). Por otra parte, se aprecian obstáculos de carácter informático para alcanzar el significado de los documentos, generados, en parte, por la relativa ineficacia de los métodos empleados por los motores de búsqueda comerciales en tareas tan complejas. Así Google o Yahoo solo abarcan del 10 al 20% de la masa documental de la Web, además de basar sus investigaciones en cadenas de caracteres y no en conceptos. Así, al buscar la palabra “gato”, es tratada como la sucesión de caracteres “g, a, t, o” y no como un concepto traducible (cat, chat,...), que pertenece a la subclase de los mamíferos, animales domésticos.

En fin principal de esta reflexión consiste en entender qué permanece y qué ha cambiado en la representación de los conceptos documentales mediante lenguajes derivados del natural en un entorno determinado por lo digital e hipertextual.

2. Estado de la cuestión

Entre los siglos XVI y XX las Ciencias naturales fueron consiguiendo un sistema de coordenadas y unidades de medida universales. La aspiración a un lenguaje unívoco y homogéneo, de aceptación y aplicación universales, se planteó desde unas premisas ajenas al lenguaje natural de uso común y exigió una desmedida actividad normalizadora para garantizar una comunicación inequívoca. La comunidad científica aceptó la lógica conceptual clásica que subyacía a los planteamientos neopositivistas por los que se usa un metalenguaje (conjunto de instrumentos simbólicos y conceptuales independientes de las lenguas naturales) para “conseguir una comunicación inequívoca y sin ambigüedad en los temas especializados” (Wüster, 1998), que se hacía mediante un lenguaje:

- altamente formalizado
- lógicamente coherente
- ampliamente compartido.

De ahí que se acercasen al léxico como si se tratase sólo de una nomenclatura o taxonomía para buscar, ante todo, objetividad, claridad y precisión lingüísticas. Las teorías pueden abundar y divergir, pero disponer de un metalenguaje común permite el diálogo, las pruebas controladas y la acumulación estructurada de innovaciones. De forma que las ciencias naturales consiguieron que una parte de sus conocimientos fueran explícitos, participables y operativos. Al centrarse en la regulación metódica y en la capacidad clasificadora de los

lenguajes, se postergó la función principal de actuar articulando la comunicación. De todas formas, si esto puede aplicarse para formalizar los lenguajes propios de las ciencias básicas, naturales y politécnicas, no sucede así cuando se consideran los rasgos del lenguaje utilizado en muchos de los dominios sociales o humanísticos, que propenden a mezclarse con el lenguaje natural común y cuyo léxico no es fácilmente jerarquizable.

Concepto de información positivista
Formas apriorísticas de representación basadas en categorías universales
Predominio de las relaciones taxonómicas
Documento como soporte del conocimiento
Información vista desde la producción o desde la recepción
Función pedagógica de la Documentación
Sistemas de información de carácter explicativo
Difusión de afán global
Búsqueda prioritaria de la precisión del lenguaje
Lenguaje depurado de las imperfecciones de los discursos de las Humanidades
El modelo eran las ciencias mas formalizadas
Sustantivo como forma de representación privilegiada
Tecnología cuyas reglas debe de observar el usuario

En la actualidad, esas 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. 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. 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.

No puede continuarse nuestra argumentación sin volver atrás y detenernos un momento en las conceptualizaciones aparecidas en 1876 cuando se dieron a conocer la *Clasificación decimal de Dewey* (Dewey, 1979) y las *Rules for a dictionary catalog* de Cutter (Cutter, 1962). La primera incluía la totalidad del conocimiento dentro de unas divisiones decimales, siguiendo la idea de aspirar a una expresión universalmente válida que diese una visión general de los conocimientos. Con características de simplicidad, ingenuidad y adaptabilidad, marcó el camino a seguir por los sistemas clasificatorios, precoordinados y de estructura jerárquica. Mientras que las teorías de Cutter siguen vigentes a través de los *Encabezamientos de materia*, de carácter precoordinado, estructura asociativa y control de vocabulario de aplicación específica a los conceptos a indizar. Sin duda la amigabilidad de uso para el usuario, frente a la rigidez arbórea de los sistemas clasificatorios,

dio a los encabezamientos una proyección hacia los lenguajes controlados superior a la que tendrían las clasificaciones.

Una causa definitiva se originó en la desmedida acumulación de información durante la Segunda Guerra Mundial. Vannevar Bush recogió su experiencia al frente del *Office for Scientific Research and Development* en el artículo "*As we may think*", donde fijó los argumentos para entender lo que varias décadas después sería conocida como *Sociedad de la Información* (Bush, 1945). Bush comprendió que la estructura secuencial de los documentos -reflejo de la secuencialidad del discurso hablado- era la causante de que los métodos de su tiempo (taxonomías alfabéticas o numéricas que disponían los conceptos en clase-subclase) fueran incapaces de procesar adecuadamente grandes cantidades de información, pues la mente humana no trabaja de esa manera, sino que opera por medio de asociaciones. Para Bush el principal problema se centraba en la forma inadecuada de almacenar y clasificar la información, por lo que propuso un sistema de procesamiento más efectivo, el *Memex*, con el que se adelantó a la importancia que tendría para los sistemas de información la recuperación mediante combinaciones lógicas correspondientes con las materias de los documentos. Daba así por superadas las jerarquizaciones y abogaba por la asociación de conceptos, imitando el modo en que las personas piensan. La principal novedad se situó en la posibilidad de indizar por asociación y de agrupar distintos tipos de información en un mismo sistema.

2.1 Los tesauros como hito

La aparición de las bases de datos durante la década de 1960 hizo que los tesauros consiguieran un enorme protagonismo en la recuperación de información. Los sistemas tradicionales eran incapaces de suministrar las soluciones necesarias para representar los conceptos contenidos en los documentos, así como las relaciones existentes entre los conceptos. Se precisaba un lenguaje estandarizado con una estructura sintáctica más simplificada que la del lenguaje natural, que controlase los sinónimos y que fuese capaz de garantizar, por inferencia, la representación del contenido informativo de los documentos de forma normalizada.

En su funcionamiento, el tesoro documental es un vocabulario destinado exclusivamente a la indización y recuperación de la información. La matriz de su vocabulario se basa en la reducción del número de términos mediante la elección de los preferentes o descriptores. Su estructura se categoriza desde unas clases generales autoexcluyentes, condición necesaria para la organización de los términos y para el establecimiento de estrategias de búsqueda deductivas, de modo que los taxones no contemplan conjuntos o niveles interseccionados. Las categorías designan aspectos particulares de un determinado dominio, permitiendo el agrupamiento de términos. De esta manera, el tesoro debe partir de una categorización y, por lo tanto, de una taxonomía del conocimiento temático. Las taxonomías están presentes en todos los *esquemas*, *tesauros*, *modelos conceptuales* y *ontologías* (Daconta, 2003), pues todos ellos contienen una clasificación de sus descriptores, relacionándolos de forma jerárquica por generalización-especialización, y estableciendo así una semántica simple.

Las asociaciones entre descriptores contribuyen a que los usuarios puedan navegar inductivamente a través de los tesauros, en cuya eficacia intervienen también:

- ◆ La integración de descriptores dentro de una misma categoría.
- ◆ El establecimiento de diferencias de unos conceptos con otros.
- ◆ La presencia de definiciones (*Scope note*), para ajustar el significado de algún descriptor.
- ◆ Cuando se construye un tesoro se manejan tres herramientas (ISO, 1986):

- Un corpus de términos extraídos del dominio cuya representación se aborda.

- Una organización general que esquematiza y segmenta ese corpus, con:

- ◆ Macroestructura global (dominio de aplicación);
- ◆ Macroestructuras secundarias (macrodescriptores que encabezan cada una de las subdivisiones del tesoro);
- ◆ Desarrollo en submacrodescriptores (clasifican las familias de términos).

- Unas microestructuras o estructuras de superficie: los propios descriptores y sus relaciones.

La funcionalidad de los tesauros les ha valido para tener una aplicación muy ventajosa, pues a su eficacia probada añaden su creación y gestión sencillas, tienen una notable coherencia y son un buen punto de partida para crear ontologías, lo que explica su número tan abundante. Pero la ambigüedad, riqueza y capacidad de innovación constante de los lenguajes en los que se encuentran expresados los documentos actuales produce en ocasiones falta de pertinencia. A lo que se añade el problema del número de documentos que circulan por la red y la variedad de sus soportes. De ahí que su uso no esté libre de mostrar algunos inconvenientes:

- Hasta la llegada de *Simple Knowledge Organisation System* (SKOS) no han dispuesto de mecanismos para compartir información en la red. Están pensados para dominios restringidos.
- En las jerarquías incluyen instancias, atributos y meronimias (palabras cuyo significado constituye una parte del significado total de otra palabra).
- Difícil adición de relaciones.
- Las reglas de nominación siguen el estándar, pero su automatización es complicada.
- Se trata de un vocabulario, sin mecanismos directos que lo relacionen con los objetos.
- No se atiene a axiomas, ni a reglas de coherencia y validación.
- Generados por consenso previo de los creadores, no de los usuarios.
- Coste elevado en la creación, mantenimiento y funcionamiento, evitables sólo mediante automatización o semiautomatización.

2.2 Las ontologías como aspiración

La gran presencia de las ontologías en la literatura profesional y académica actuales se origina al apreciarlas como uno de los recursos significativos para la Web Semántica, pues resulta evidente la necesidad de modelar semánticamente los conceptos en un sistema de organización del conocimiento (Antonioni, 2004: 10-18). Estamos además ante un término en boga, cuyo resalte proviene de avalar el requisito de descripción semántica de una materia en el entorno Web para que pueda ser interpretada por las aplicaciones informáticas de manera comprensible para los usuarios. Al aplicarse las ontologías en y por medio de Internet han tenido que habilitar elementos para mejorar su interoperabilidad y reutilización, y han adaptado su expresión a lenguajes Web (XML o basados en XML). Funcionalmente, han heredado los frutos de las investigaciones hechas en los Sistemas de Organización del Conocimiento o KOS -*Knowledge Organization Systems*- ahora comprendidos dentro de las ontologías (Zeng y Chan, 2004). De manera que una de las causas que han llevado al uso de las ontologías en diferentes disciplinas está, así, en que ya existía una forma de representar el conocimiento de un área, pero que hasta ahora recibía otra denominación. Por otra parte, las ontologías se constituyen indudablemente en un ingenio práctico para muchos sistemas de los que se espera que dispongan de una estructura que relacione sus elementos y que cuente con cierto grado de inferencia o razonamiento.

Las ontologías tienen como misión representar el conocimiento a partir de la organización taxonómica en cuanto modo de clasificación o categorización jerárquica de los conceptos pertenecientes a un conjunto temático, siguiendo, por lo general, una configuración arborescente que establece entre los conceptos que integra una relación de generalización-especialización, es decir, asociando los términos por *subclase-de* o *subclasificación*-

de. De este modo aplican una semántica simple de acuerdo con algunas de las propiedades que caracterizan a las ontologías (Daconta et al., 2003: 145; Z3919:2005: 9). Por este motivo, se trata de clasificar o categorizar un conjunto de conceptos con semántica más compleja que varían desde taxonomías a las ontologías más completas y formalizadas.

Una ontología es una organización cognitiva que conforma un sistema de organización del conocimiento. Sin embargo, uno de los principales problemas para representar el conocimiento es el consenso sobre qué representar y cómo hacerlo, cuestión que se ha abordado con diferentes modelos desde diferentes disciplinas: Biblioteconomía y Documentación, Inteligencia Artificial, Ingeniería del Software, Lingüística, Ingeniería Ontológica, etc. De forma que el grado de representación semántica y la finalidad buscada condicionan los modelos y lenguajes a la hora de construir un sistema de organización del conocimiento englobable dentro del espectro de las ontologías. Se puede considerar, pues, como ontología cualquier representación genérica o concreta de información desde la noción más simple de una taxonomía, pasando por los tesauros y modelos conceptuales hasta llegar a las teorías lógicas o noción más compleja. En cuanto sistema de organización, las ontologías pueden ser configuradas siguiendo distintas técnicas de modelado del conocimiento y pueden ser puestas en funcionamiento con diversos lenguajes formales. Por ello tanto, su concepto sería más clarificador y completo si lo situásemos en un contexto que no olvidase la finalidad para la que se construye el recurso.

El potencial de las ontologías dentro de la Web Semántica viene determinado por la idoneidad desarrollada para favorecer la interoperabilidad y la reusabilidad, que se fundamentan en:

- La adopción de lenguajes comunes y compatibles (con un lenguaje de sintaxis XML, y la expresión normalizada del conocimiento mediante las tripletas recurso-atributo-valor, esto es RDF).
- La referencia a vocabularios de metadatos para desambiguar conceptos (p.e: DC, SKOS).
- La creación de ontologías con capacidad de reutilización, como las de alto nivel (*Top Ontologies*) y las de amplio uso (*Generic Ontologies*).
- La adopción de paradigmas comunes para expresar el conocimiento (p.e.: OWL DL o SCL).

Casi todas las representaciones cognitivas que veremos luego (*Tesauros automáticos o conceptuales, tesauros de verbos, Topic maps*) usan ontologías en su funcionamiento. Para mejorar la precisión de las recuperaciones documentales aprovechan el diseño de ontologías por áreas del conocimiento desde las que se autogeneran tesauros que permitan distinguir los sinónimos, suprimir los homónimos e inducir relaciones asociativas entre los descriptores. La aspiración consiste en alcanzar ontologías que abarquen los diferentes tipos de documentos, las descripciones conceptuales, las relaciones entre dichos documentos (citas), y las de éstos con los diferentes problemas científicos; además de índices, descripciones bibliográficas, tesauros, códigos clasificatorios, información terminológica, etc. Su aplicación debe proporcionar una visión general de la estructura y de la terminología del dominio que facilite recuperaciones relevantes. Sintetizando, para una comprensión clara de lo que sea una ontología aplicada a los lenguajes documentales, se trata de una descripción formal de los conceptos y de las relaciones que se dan entre los conceptos (Gruber, 1992).

3. Tendencias en la representación mediante lenguajes documentales.

A partir de los años noventa los presupuestos de la postmodernidad han hecho que los tesauros se activen con nuevas relaciones identificadas, y en mayor número. Han aparecido nuevas propuestas de navegación y de visualización mediante grafos explícitos de conexiones informativas. Podemos decir que Internet y su oferta de enlace hipertextual de documentos ha obligado a diferenciar la representación de los contenidos documentales, pasando a buscarse soluciones a través, primero, de los Tesauros automáticos o con la intervención de

Tesauros verbales, luego con las búsqueda de mapas de conceptos que acabó en la metodología mixta de los *Topic maps*.

Concepto postmoderno de información
Categorías derivadas del carácter funcional
Predominio de las relaciones asociativas
Primacía de la función comunicativa
Información vista desde la relación producción-recepción del n
El usuario como sujeto de la interpretación
Sistemas incluidos en el proceso de mediación
Relativización por el contexto o situación
El lenguaje visto desde su función comunicativa
Utilización de toda la riqueza expresiva del lenguaje
Cada documento contiene un modelo léxico
Léxico más aproximado al lenguaje natural
Tecnología al servicio de la creatividad

Los *Tesauros automáticos o conceptuales* conforman una red semántica en la que cada nodo contiene un único concepto semántico que puede tener una serie de descriptores asociados que se identificarían según las relaciones habituales en los tesauros: preferenciales, jerárquicas o asociativas. En su funcionamiento es destacable un incremento de las relaciones, en especial las de asociación, las que más subtipos y subdivisiones presentan, lo que ha crecido el número posible de relaciones asociativas presentes en un tesoro (Tudhope, 2001):

- ◆ Ideas combinadas.
- ◆ Términos relacionados conceptualmente.
- ◆ Contigüidad.
- ◆ Relaciones asociativas por definición.
- ◆ Relaciones asociativas con diferente jerarquía.
- ◆ Relaciones asociativas trasladadas por significado.
- ◆ Cuestiones de finalidad.
- ◆ Otras relaciones asociativas sin especificar.

La denominación de tesoro automático o conceptual se fundamenta en la noción de materia (concepto) de la que tratan los textos, que aúna términos y conceptos por similitud de su sentido desde el contexto del usuario, y que se distingue precisamente por su riqueza en relaciones asociativas. Puede verse como una *Red semántica conceptual* por la que se navega desde los términos más genéricos de una faceta hacia los más específicos, e inversamente (navegación vertical), que a su vez permite la transición de una clase hacia otra y de un campo de la ciencia hacia otro mediante las relaciones asociativas (navegación horizontal por nudos polijerárquicos). E incluso como un *Espacio conceptual* donde el tesoro aparece como un sistema formal definido por un dominio algebraico. Su modelo *espacial* define las relaciones entre términos con mayor precisión que los tesauros tradicionales. Respecto a los tesauros convencionales presentan estas novedades:

- a) Listan todas las palabras “no vacías” existentes en las bases de datos,
- b) consideran los términos coloquiales, incluso las variaciones y truncamientos de los términos reconocidos (los *Topic Maps* permitirán acceder a un concepto por todos sus sinónimos, incluso en siglas y códigos).
- c) aportan notas definitorias que despejen las posibles dudas de uso.
- d) razonan las equivalencias existentes entre términos.
- e) contienen numerosas relaciones asociativas, incluso con los no descriptores.

En los tesauros conceptuales los términos se extraen de documentos a texto completo, para luego conformar las bases de conocimiento descentralizadas de Internet. Los enlaces en la red se establecen tras su adaptación al espacio conceptual hipertextual mediante el lenguaje HTML o el XML. Se obtiene un corpus terminológico cuya representación se establece como una red semántica neuronal: en cada nodo hay un concepto semántico con el que se asocian una serie de descriptores. Los enlaces entre los descriptores pueden también establecerse según las típicas relaciones de equivalencia, facetadas o por asociación, mientras que la recuperación se establece desde la pregunta del usuario. Los conceptos que este ha solicitado se confrontan con la red terminológica, cuyos elementos están diseñados como mapas representativos de los textos, y que actúan así como lenguajes controlados que organizan la información de cualquier objeto disponible en la red. De todas formas, se observan también algunas limitaciones:

- ◆ *Desorientación de los usuarios* para encontrar los conceptos apropiados durante su navegación por los textos, debida a la saturación de enlaces en los nodos que conjuntan información de numerosas fuentes.
- ◆ Deficiente normalización terminológica de los documentos, soslayable mediante bases de corpus terminológicos de cada una de las lenguas más usadas.
- ◆ *Las interfaces hombre-máquina deben ser capaces de relacionar los conceptos a partir de ontologías preestablecidas.* El acceso se allanaría con representaciones gráficas de la red (tabla de contenidos gráficos con las conexiones visibles mediante enlaces hipertextuales: mapas).

Otra de las propuestas de mejora de los tesauros es la inclusión de verbos que complementen a los estáticos tesauros tradicionales de sustantivos (Levin, 1993). El uso de descriptores verbales aporta múltiples ventajas, como la posibilidad de indizar audiovisuales mediante gerundios, la identificación verbal de asociaciones funcionales mucho más adaptables a dominios concretos, la posibilidad de mostrar la relación existente entre dos conceptos usando los inmensos medios del lenguaje natural (categorías verbales a modo de relaciones facetables), y la desambiguación conceptual.

La tipología de relaciones de asociación del lenguaje UML se integra así en la tendencia a ampliar el número de relaciones de asociación de los tesauros para que no planteen ninguna ambigüedad, como la relación de agregación, en la que la desaparición del todo no implica la desaparición de las partes; De composición, en la que la desaparición del todo implica la desaparición de sus partes; Información de multiplicidad (es decir, cuántos objetos pueden interactuar en una misma relación); Dirección de la relación; y Tipificación de relaciones.

Esta aproximación mediante la integración verbal procede del área pedagógica, donde esta forma de relacionar los conceptos mediante verbos recibió la denominación de mapas conceptuales (*concept maps*).

Es directa la aplicación a la automatización en la construcción de tesauros por la aparición de sustantivos en proximidad a estructuras verbales, siguiendo un proceso que comienza con el análisis de documentos relevantes para la extracción de su vocabulario (glosarios, diccionarios, estándares sobre vocabulario, etc), luego se procede a una depuración manual del vocabulario extraído, obteniendo los descriptores, con los que se indizan los documentos textuales maestros (manuales, estándares, artículos, ...). En esta etapa se almacenan principalmente aquellas frases del documento en las que aparece uno o varios descriptores del tesoro en el Sintagma Nominal Sujeto, y uno o varios descriptores en el Sintagma Verbal. Posteriormente, los conceptos dinámicos se agrupan, clasifican y se asimilan a las relaciones del tesoro que se deseen identificar. Finalmente, se revisan a mano en el tesoro las relaciones obtenidas.

Otro caso que ha supuesto innovación en la representación de la información se ha originado en los Mapas conceptuales de navegación por redes semánticas. Su estudio se originó en la necesidad de crear índices en torno a algún concepto o materia. Las redes semánticas son un método común de representar el conocimiento en el campo de la inteligencia artificial, que busca establecer comunicación entre las personas y las máquinas:

- Grafos o redes conceptuales constituidos por conceptos y relaciones de conceptos.

- Son colecciones ordenadas de nodos conectados por arcos que se usan para representar documentos.
- Un tipo de grafo es la red semántica, que representa las relaciones semánticas que se establecen en el texto.
- No aplican ningún control de términos

Los mapas conceptuales ofrecen una red de relaciones más rica que los tesauros. El concepto que origina su estructura (nodo-enlace-nodo) es próximo a su equivalente en las redes hipertextuales por lo que soportan la navegación de un modo muy natural. El uso de mapas conceptuales permite desarrollar mejores mecanismos de representación y recuperación, ya que las relaciones entre los conceptos se eligen teniendo en cuenta las necesidades y expectativas de cada usuario, siguiendo los pasos: 1. Selección de los conceptos que se representarán en el mapa. 2. Listado de esos conceptos. 3. Agrupación de los conceptos relacionados. 4. Ordenación de los mismos en forma bidimensional o tridimensional. 5. Enlace de cada par de conceptos mediante líneas etiquetadas.

Se trata de una técnica para representar el conocimiento en gráficas cognitivas que establecen redes de conceptos que se componen de **nodos** (puntos / vértices) que representan conceptos y de **enlaces** (*arcs*: arcos / *edges*: extremos, satélites) que representan las relaciones entre los conceptos. De ellos surgió la idea de establecer un nuevo estándar conocido como *Topic map*: un documento, o un conjunto de documentos SGML o XML interrelacionados en un espacio multidimensional en el que las localizaciones son *Topic* (ISO, 2000). Como elementos de un *Topic Map* hay que enumerar: **Topic** p.e. <CALSI>; **Topic type** p.e. <jornadas científicas>; **Association** <tiene lugar en>; **Association type** <tener lugar en> (localización); **Scope** (ámbito en el que una relación tiene sentido) **theme** <Sociedad del conocimiento>; **Topic occurrence** (<http://www.calsi.com>); Occurrence type (p.e. Página web).

Los *Topic Maps* presentan indudables ventajas, como haber optimizado los mapas conceptuales y la fusión de vocabularios jerarquizados o no. Además, es un estándar ISO intuitivo tanto en su creación como en su interpretación. Es, junto con RDF/OWL y UML, uno de los lenguajes más difundidos en la Web Semántica, óptimo para desarrollar portales y para la expansión de búsquedas. Por el contrario, como desventajas suyas cuentan que no tienen inferencia, reglas ni axiomas. Son poco flexibles cuando se aplican a otros entornos, p.e. ingeniería del software. Compiten con desventaja frente a RDF/OWL.

3. Permanencia de las palabras-clave 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 el espacio Web compartido y abierto. Propuso este neologismo Thomas van der Wal al fusionar 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 muestran mucho 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.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):

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

Clasificación de las folksonomías según Hammond (2005)

4.1 Ventajas de las folksonomías

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

- ◆ **Simplicidad de 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.
- ◆ **Adecuación al entorno Web:** es la única solución posible para indizar los enormes volúmenes de información en la red, sobre todo cuando la información a indizar no es textual, exigiendo una indización manual, como en el caso 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. Además permite recuperar, como ya se ha comentado, material multimedia (solucionando parte del 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.

4.2 Desventajas

Las folksonomías presentan asimismo notables inconvenientes, lo cual las lleva a ser un instrumento útil en materiales no excesivamente críticos. Los principales inconvenientes se derivan de (Al-Khalifa, 2007):

- ◆ 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.

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. 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 tesoro 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. Las clasificaciones propuestas por los filósofos clásicos cuando se acercaban a poner las bases retóricas de los discursos 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 asignación a los contenidos digitales de palabras libres por los propios usuarios, 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 productiva.

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 sí.

Lamentablemente, la Web Semántica, operativa desde 1999, no ha tenido el éxito esperado. Las causas son diversas, pero debemos considerar entre ellas:

- ◆ Falta de legibilidad de los lenguajes RDF y OWL, que supone un cuello de botella para que los expertos validen las ontologías 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 efectos como usabilidad, legibilidad o necesidad de formación previa para su interpretación.
- ◆ 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
- ◆ La migración de folksonomías a folkontologías está aún por desarrollar, aunque ya existen proyectos que lo intentan.
- ◆ Falta incorporar técnicas semiautomáticas para la creación de Sistemas de Organización del Conocimiento basadas en PLN y Minería de Datos, ya que la carencia o la lentitud de creación de estos recursos dificulta la implantación de la Web Semántica.
- ◆ Presencia de duplicidades en los Vocabularios de Metadatos y Ontologías, lo que provoca al usuario desconfianza y confusión pues 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 del W3C, los PSI de los Topic Maps, Zthes y MADS.

Referencias

- (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
- (Antoniou, 2004) Antoniou, G. y Harmelen, F. van.- *A Semantic Web Primer*. London: The MIT Press, 2004.
- (Bush, 1945) Bush, V.- As we may think, en *Atlantic Monthly*, 1945, 176: 101-108.
- (Daconta et al., 2003) Daconta, Michael C.; Obrst, Leo J. y Smith, Kevin T.- *The Semantic Web. A guide to the future of XML, Web Services, and Knowledge Management*. Indianapolis: Wiley, 2003.
- (Dewey, 1979) Dewey, M.- *Decimal classification and relative index*. 19th ed. Albany (New York): Forest Press, 1979. 3 v.
- (Hammond, 2005) Hammond, T., T. Hannay, B. Lund and J. Scott.- Social Bookmarking Tools (I): A General Review, en *D-Lib Magazine*, 2005, 11, nº 4: 05.
- (ISO, 1986) International Organization for Standardization, ISO 2788:1986. *Guidelines for the establishment and development of monolingual thesauri*. Geneva: ISO, 1986
- (Cutter, 1962) Cutter, Ch.- *Rules for a dictionary catalog*. 4th ed. London: Chaucer House, Malet Place, 1962.
- (Gruber, 1992) Gruber, Thomas.- *Toward principles for the design of ontologies used for knowledge sharing*, en Guarino, Nicola y Poli, Roberto (Eds.).- *Formal Ontology in Conceptual Analysis and Knowledge Representation*. Boston: Kluwer Academic Publishers. Paper presented at the International Workshop on Formal Ontology, March, 1993, Padova.
- (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.
- (ISO, 2000) *ISO/IEC 13250: 2000. SGML-Topic Maps*.
- (Levin, 1993) Levin, B.- *English Verb Classes and Alternations: A preliminary Investigation*. Chicago: The University of Chicago Press. 1993.
- (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.

(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].

(Wüster, 1998) Wüster, E.- *Introducción a la teoría general de la terminología y a la lexicografía terminológica*. Barcelona: Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada (IULA), 1998: 48.

(Z3919:2005) ISO. ANSI/NISO. Z39.19-2005. Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. [En línea] <http://www.niso.org/standards/index.html>

(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 2007) 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.