Taxonomías para la categorización y la organización de la información en sitios web

Miquel Centelles

Citación recomendada: Miquel Centelles. *Taxonomías para la categorización y la organización de la información en sitios web* [en linea]. "Hipertext.net", núm. 3, 2005. <http://www.hipertext.net> [Consulta: 30 ene. 2007]. .

1. Concepto de taxonomía
2. Construcción de la taxonomía
2.1. Automatización de los procesos de construcción de la taxonomía
3. Categorización de recursos
4. Aplicación de la taxonomía en el desarrollo de sistemas de búsqueda de información
5. Bibliografía
6. Notas

1. Concepto de taxonomía

En el momento en que se publique este artículo, se habrá producido un hecho que ha de marcar un antes y un después en la evolución de las taxonomías como sistemas de organización de contenidos: la aparición del borrador final de la revisión de la norma ANSI/NISO Z39.19-1993, *Guidelines for the construction, format, and management of monolingual thesauri* [1] . Esta revisión ha sido llevada a cabo entre 2002 y 2004 por el Thesaurus Advisory Group (desde ahora, TAG), creado en el seno de la National Information Standards Organization, y que persigue la introducción de un lenguaje más amigable en la norma, la actualización de su contenido al entorno actual de la información digital, y la ampliación de su alcance a la variada gama de organizaciones productoras y de contenidos.

No disponemos de un borrador de la norma revisada pero sí de un sumario de su contenido y de las notas de las reuniones llevadas a cabo por el TAG. A partir de estos documentos, se puede observar como una de las modificaciones globales que se han propuesto es cambiar el título de la norma " *Guidelines for the construction, format, and management of monolingual thesauri* " por el de *Construction, format and management of monolingual controlled vocabularies* . Los vocabularios controlados incluyen cuatro tipos principales: las listas, los anillos de sinónimos, las taxonomías y los tesauros. La revisión de la norma ANSI/NISO Z39.19 se propone definir "normativamente" los cuatro tipos, y establecer los elementos fundamentales de construcción y gestión de todos ellos. Concretamente, en la "TAG Conference Call, June 30, 2003" (2003), se incluye las definiciones provisionales que reproducimos a continuación:

> - Lista: "A set of words or phrases displayed in an organized series."
>
> - Anillo de sinónimos: "A set of words or phrases that are consireded to be equivalent for the purposes of retrieval. Synonym rings are not used during input."
>
> - Taxonomía: "An organized set of words or phrases used for organizing information and primarily intended for browsing."
>
> - Tesauro: "A controlled vocabulary that indicates preferred terms, variant terms, and term relationship. Usually considered to be the most complex of controlled vocabularies." A partir de las modificaciones propuestas por el TAG, la definición definitiva es la siguiente: "A set of words or phrases with equivalent terms explicitly identified and with ambiguous words or phrases (e.g. homographs) made unique. This set of terms also may include broader-narrower or other relationships."

De acuerdo con esta definición, la taxonomía no exige que sus componentes estén conectados mediante un tipo específico de relaciones; simplemente requiere que sus componentes estén organizados. Las características definitorias son su finalidad "prioriza la exploración ("browsing")" y, por lo tanto, su entorno de aplicación "el entorno digital".

En cambio, en algunos documentos relativos al proceso de revisión de la norma ANSI/NISO Z39.19 la diferencia entre los cuatro tipos de vocabularios controlados está determinada por la menor o mayor complejidad estructural que presentan. En un extremo, las listas y los anillos de sinónimos se limitan a incorporar la relación de equivalencia; en el otro extremo, los tesaurus incorporan las relaciones de equivalencia, jerarquía y asociativa. En una posición central, las taxonomías incorporan las relaciones de equivalencia y de jerarquía.

En espera de que los trabajos del TAG aporte una definición normativa del concepto de taxonomía, debemos destacar que, en la actualidad, no disponemos de un concepto universalmente aceptado de dicho término.

Etimológicamente hablando, taxonomía procede de los términos griegos "taxis" , ordenación, y "nomos", norma. Aristóteles fue uno de los primeros en utilizar este término, en el 300 antes de Cristo, para designar e squemas jerárquicos orientados a la clasificación de objetos científicos. El botánico Carlos Linneo (1707-1778) designó con el término taxonomía a la clasificación de los seres vivos en agrupaciones jerárquicamente ordenadas de más genéricas a más específicas (reino, clase, orden, género, y especies). A partir de esta concepción clásica, se desarrolló la taxonomía

como un subcampo de la biología dedicado a la clasificación de organismos de acuerdo con sus diferencias y similitudes. De acuerdo con Grove (2003, p. 2774), los principios que proporcionaban una guía rigurosa para la construcción de taxonomías eran la base lógica, la observación empírica, la estructura jerárquica basada en la herencia de propiedades, la historia evolutiva, y la utilidad pragmática. Las fuentes terminológicas de la lengua general todavía recogen el significado especialmente orientado al entorno de las ciencias experimentales, como demuestra el artículo que incorpora la última edición en papel del *Diccionario de la lengua española* (2001):

"1. f. Ciencia que trata de los principios, métodos y fines de la clasificación. Se aplica en particular, dentro de la biología, para la ordenación jerarquizada y sistemática, con sus nombres, de los grupos de animales y de vegetales.

2. f. clasificación (acción y efecto de clasificar)."

En su concepción clásica, vinculada a las ciencias experimentales, la taxonomía aplica un criterio monojerárquico en el establecimiento de los sistemas de clasificación; es decir: cada una de las agrupaciones o clases que lo componen sólo puede ocupar un lugar, y sólo uno, en la estructura jerárquica.

A principios de los años 90 del siglo XX, el concepto de taxonomía se incorpora a otros ámbitos del conocimiento, como la psicología, las ciencias sociales y la informática, para designar casi todos los sistemas de acceso a la información que intentan establecer coincidencias entre la terminología del usuario y del sistema. Los primeros especialistas que desarrollaron sistemas de organización de contenidos para la Web formaban parte del área de consultoría en gestión del conocimiento, y procedían de ámbitos próximos a la informática y la ingeniería (gestión de contenidos y arquitectura de la información); no conociendo la tradición de los lenguaje documentales del ámbito de la Documentación, asignaron el término taxonomía para los sistemas que desarrollaban. Este término se mantiene en uso actualmente para designar los sistemas de organización de contenidos en el contexto de Internet, aunque la teoría y la práctica de los lenguajes documentales se ha venido aplicando de forma intensiva en este contexto.

Antes de proponer una definición del término taxonomía, acorde con los ámbitos de desarrollo actuales, hemos realizado un trabajo de identificación y confrontación de los rasgos semánticos con que se define. Para ello, hemos realizado una amplia búsqueda de definiciones en todos los ámbitos de estudio, desarrollo y/o aplicación del término taxonomía. A priori, no hemos impuesto limitación alguna al origen de las definiciones; únicamente hemos descartado aquéllas elaboradas a partir de una concepción clásica del término. El resultado ha sido la localización de 36 definiciones publicadas en el período comprendido entre 2000 y 2005 en diferentes tipos de fuentes [2] :

- Diccionarios y enciclopedias especializadas.
- Artículos de consultores en gestión de contenidos y arquitectura de la información.
- Artículos de especialistas vinculados al ámbito académico.
- Documentos técnicos y comerciales de aplicaciones informáticas para el desarrollo de taxonomías.

El análisis de las definiciones muestra que éstas inciden sobre cuatro variables: el lugar que ocupa la taxonomía en el ámbito de los sistemas de organización del conocimiento (en adelante, SOC); el contexto informativo en que se aplican la taxonomía; las finalidades que persigue la taxonomía; y el modelo estructural con que se interrelacionan los elementos que componen la taxonomía.

Entre las definiciones que hacen referencia al lugar que ocupa la taxonomía en el marco de los SOC (13 de 36), la opinión mayoritariamente aceptada es considerarla como un tipo de vocabulario controlado (5 de 13) o, incluso, un tipo específico de tesauro o esquema de clasificación (3 de 13). No faltan, sin embargo las opiniones que la consideran como una categoría amplia que incorpora modalidades específicas como los tesauros (4 de 13). En este caso, la taxonomía puede ser definida como el proceso general de organización o clasificación de contenidos:

"In the 1990s, taxonomy was redefined as any semantically significant, systematic organization of content or as the process of developing such an organization." (Grove 2003, p. 2.770)

o incluso puede ser elevada al rango de ciencia:

"The science of categorization, or classification, of things based on a predetermined system. In reference to Web sites and portals, a site'ss taxonomy is the way it organizes its data into categories and subcategories, sometimes displayed in a site map." (Webopedia)

Más de la mitad de las definiciones (21 de 36) restringen el ámbito de aplicación de las taxonomías; 14 de ellas, a entornos digitales y, más específicamente, al desarrollo de sitios web; 11 a entornos corporativos y, más específicamente, de empresas [3] . En cuatro casos convergen ambos criterios de restricción del significado de taxonomía; una buena muestra de esta corriente, es la definición propuesta por Gilchrist, Kibby y Mahon (2000, p. 6), que ha alcanzado un importante factor de impacto en la bibliografía especializada:

"- a correlation of the different functional languages used by the enterprise

- to support a mechanism for navigating, and gaining access to the intellectual capital of the enterprise

- by providing such tools as portal navigation aids, authority for tagging documents and other information objects, support for search engines, and knowledge maps

- and possibly, a knowledge base in its own right."

El resto de las definiciones (15 de 36) no imponen restricciones al ámbito de aplicación de la taxonomía, es decir; abarcan todos los soportes documentales, todas las áreas de conocimiento y profesionales, etc.; y tanto los entornos analógicos como digitales.

Las definiciones que vinculan las taxonomías al entorno digital destacan, como finalidades prioritarias, la mejora de la navegación y el desarrollo de sistemas de búsqueda basados en la exploración ("browsing") y en la recuperación ("searching"). Las definiciones que vinculan las taxonomías al entorno corporativo destacan el valor estratégico de las taxonomías en áreas como la gestión del capital intelectual y, en general, del conocimiento. Una muestra de una definición que otorga a la taxonomía una posición estratégica en el desarrollo de sitios web corporativos es Taxonomy strategies:

"Overall scheme for organizing content to solve a business problem such as improving search, browsing for content on an enterprise-wide portal, enabling business users to syndicate content, and otherwise providing the basis for content re-use."

Desde el punto de vista estructural, 23 de las 36 definiciones consideran que las taxonomías se caracterizan por la aplicación de la relación jerárquica entre los elementos que organiza. En los casos en que la definición de taxonomía se orienta a su posición en el marco de los vocabularios controlados (Fast, Leise y Steckel, 2003 y NISO/ANSI Z39.19) las definiciones asignan a la taxonomía una posición central determinada por la aplicación de las relaciones de equivalencia y de jerarquía. Las listas y anillos de sinónimos constituyen vocabularios controlados más simples desde el punto de vista estructural, ya que únicamente aplican la relación de equivalencia; en el otro extremo, los tesauros constituyen el máximo nivel de complejidad estructural, ya que a las relaciones de equivalencia y jerarquía incorporan la asociativa. Sólo una incorpora la relación asociativa en la definición del concepto:

"The basic idea behind taxonomy is to provide a controlled vocabulary for metadata attributes, and to specify relationships between terms in the controlled vocabulary. The simplest relationships are broader, narrower, and related, but relationships can be much more specific and complex."

Entre las 23 definiciones que privilegian la relación jerárquica en el concepto de taxonomía, seis incluyen alguna consideración sobre la monojerarquía o la polijerarquía en relación al concepto de taxonomía: dos declaran explícitamente que la monojerarquía es la relación óptima, y mantienen, por lo tanto, la perspectiva clásica de las ciencias naturales; dos admiten los dos tipos de jerarquía como posibles; y dos manifiestan la preferencia por la polijerarquía.

A partir de la documentación elaborada por el TAG de la NISO, y a la luz de las propiedades mayoritariamente aceptadas en la definiciones formuladas en los ámbitos de estudio, desarrollo y/o aplicación, proponemos la siguiente definición:

Una taxonomía es un tipo de vocabulario controlado en que todos los términos están conectados mediante algún modelo estructural (jerárquico, arbóreo, facetado...) y especialmente orientado a los sistemas de navegación, organización y búsqueda de contenidos de los sitios web.

Es preciso realizar tres puntualizaciones sobre el contenido de esta definición:

> - Los términos (o categorías) representan algún aspecto del contenido, contexto o estructura de los recursos de información, y no únicamente del contenido.
>
> - Los modelos estructurales no suelen presentarse de forma pura; es posible (y, en el mundo real, habitual) que una misma taxonomía presente estructuras resultantes de la mezcla de modelos.
>
> - Los documentos que reflejan las discusiones en el seno del TAG revelan una falta de consenso en relación a las aplicaciones y usos preferentes de las taxonomías. Algunas notas de las reuniones de dicho grupo (por ejemplo, "TAG Conference Call, may 19, 2003" (2003)), reflejan cómo inicialmente la concepción de la taxonomía se orientó a la exploración ("browsing") y a la navegación en perjuicio de la recuperación ("searching"); en la versión final de la definición de taxonomía su aplicación abarca también este último mecanismo.
>
> - Excluimos del concepto de taxonomía las folksonomías o clasificaciones distribuidas (Mathes, 2004).

Una vez establecida una definición de taxonomía, vamos a realizar un breve recorrido por los procesos de construcción de taxonomías y su aplicación en la categorización de recursos y el desarrollo de sistemas de búsqueda de información de los sitios web. Ambos procesos deben estar precedidos por una planificación estratégica que determine qué características debe presentar la taxonomía a partir del análisis del contexto "que identificará las prioridades de la corporación en la organización y presentación de la información en el sitio web", de la audiencia "que identificará las necesidades y comportamientos de búsqueda y uso de la información por parte de los diferentes segmentos de usuarios" y del contenido "que identificará patrones de contenidos".

2. Construcción de la taxonomía

La construcción de las taxonomías corporativas supone la realización de cuatro procesos:

1. Delimitación de la realidad (entidad, área de conocimiento, sector industrial, etc.) que será representada por la taxonomía.

2. Extracción del conjunto de términos o categorías que representan dicha realidad.

Para llevar a cabo este proceso es necesario establecer, en primer lugar, cuáles son las fuentes prioritarias y los mecanismos de extracción idóneos para cada una de ellas. Existen tres tipos: las fuentes personales, integradas por usuarios del web y especialistas en el dominio del web; fuentes documentales, integradas por documentos representativos de los tipos de contenidos identificados en la fase de planificación estratégica; y las taxonomías o instrumentos de representación del conocimiento ya existentes (desde nomenclaturas de las unidades y recursos existentes en una entidad a los cuadros de clasificación de la documentación administrativa).

Es necesario identificar los mecanismos de extracción para cada una de las fuentes; así, en el caso de las fuentes personales, resultan especialmente útiles las entrevistas con representantes de los usuarios del sitio web, y el análisis de los registros de transacciones de búsquedas y consultas.

El resultado de este proceso es un registro de términos o categorías representativas.

3. Control terminológico de los términos o categorías.

Este proceso supone la realización de dos tareas. En primer lugar, se identifican los diferentes términos que designan un mismo concepto; en caso de que sean dos o más es necesario determinar cuál se considera preferente y cuáles se consideran no preferentes. En segundo lugar, es necesario dar una forma correcta y consistente a todos los términos de la taxonomía, independientemente de si son preferentes o no preferentes.

El resultado de este proceso es el establecimiento de la relación de equivalencia entre todos los términos de la taxonomía.

4. Establecimiento del esquema y la estructura de organización de los términos o categorías

El esquema de organización incluye el criterio o criterios utilizados para dividir y agrupar las categorías . A priori, los criterios son ilimitados y su idoneidad depende del objeto que deba representarse mediante la taxonomía. Ejemplos de los criterios más utilizados son: los temas, las materias y/o disciplinas; las personas; las entidades; los destinatarios; los procesos, tareas y/o funciones; los tipos de documentos; etc.

El modelo estructural define el tipo de relación que se establece entre las agrupaciones de categorías derivadas del esquema de organización. La tendencia general ha sido aplicar los modelos jerárquico (basado en la relación "tipo de") y arbóreo (basado en la relación "parte de") y, de hecho, las normas internacionales y nacionales de construcción de tesauros que se han aplicado a las taxonomías corporativas encumbran estos dos modelos estructurales. Un tercer modelo, el facetado, constituye una buena alternativa para el entorno hipertextual, en que resulta clave la descomposición de las diferentes perspectivas desde las que se puede observar un mismo concepto o ítem. De hecho, este modelo se está utilizando cada vez más para determinados tipos de sitio web No obstante, la documentación de que disponemos sobre la revisión de la Norma ANSI/NISO Z39.19 no parece que vaya a incorporar esta alternativa.

Tradicionalmente, se han distinguido dos técnicas para el desarrollo de la estructura de la taxonomía: la técnica de arriba a abajo ("up to down") y la técnica de abajo a arriba ("down to up").

- La aplicación de la técnica de arriba a abajo supone la identificación inicial de un número limitado de categorías superiores, y la agrupación del resto de categorías en niveles sucesivos de subordinación hasta alcanzar los niveles de categorías más específicas. Esta técnica puede orientarse tanto a la aplicación de un modelo estructural jerárquico (y/o arbóreo) como facetado. La posibilidad de ejercer un control previo sobre las categorías principales hace que esta técnica se aplique a la construcción de taxonomías que tienen, como finalidad exclusiva o prioritaria, el desarrollo de sistemas de exploración ("browsing") y/o navegación.

- La aplicación de la técnica de abajo a arriba se basa en la identificación inicial de las categorías más específicas, que van agrupándose en niveles sucesivos de superordinación hasta alcanzar el nivel de categorías superiores. Generalmente, esta técnica se ha orientado, fundamentalmente, a la aplicación de un modelo estructural jerárquico (y/o arbóreo), aunque, como en el caso anterior, puede facilitar el análisis para la toma de una decisión sobre el modelo estructural que resulta idóneo aplicar. En cualquier caso, es la técnica que se ha aplicado al desarrollo de métodos de intervención de representantes de los usuarios reales y potenciales en el establecimiento de la estructura de las taxonomías (por ejemplo, el método de la ordenación de fichas o "card sorting").

2.1. Automatización de los procesos de construcción de la taxonomía

Un factor crítico en la construcción de la taxonomía es el grado de automatización que se aplica a los procesos anteriormente indicados. El grado de automatización puede contemplarse como un *continuum* : en un extremo se sitúan los sistemas manuales (o intelectuales), y, en el otro, los automáticos. En un punto central, se sitúan los sistemas semiautomáticos.

Cabe destacar que en el momento actual difícilmente se aplican sistemas completamente manuales para la creación de taxonomías.

En el nivel mínimo de automatización encontramos dos tipos de soluciones: las taxonomías preelaboradas ("taxonomy templates"), especializadas en un sector industrial determinado, que deben ser adaptadas a las condiciones específicas de una organización determinada [4] , y las herramientas de edición de taxonomías. Este segundo tipo de soluciones ofrecen a los administradores de la taxonomía un depósito para la gestión de términos, un entorno amigable para el establecimiento de relaciones entre los términos, y diferentes modalidades de presentación y visualización de los resultados. Muchas de estas aplicaciones ya existían como administradores de tesauros, y no han incorporado excesivas innovaciones para su nuevo cometido en el contexto de las taxonomías. Como ejemplos de estas modalidades podemos citar los productos Multites 2005 (http://www.multites.com) o Term Tree (http://www.termtree.com.au).

En el nivel máximo de automatización, encontramos programas que analizan el corpus de recursos digitales de un sitio web, y extraen categorías mejor dicho, agrupaciones de recursos ("clusters") mediante la aplicación de análisis estadístico y/o procesamiento lingüístico. Generalmente, el proceso de construcción de la taxonomía y de categorización de los recursos es simultáneo; incluso en algunos casos, el resultado es directamente editable como sistema de exploración ("browsing"). Una opción extrema de esta modalidad de automatización es la que da lugar a las denominadas taxonomías dinámicas: agrupaciones de recursos resultantes de una consulta a un buscador que suele responder más a análisis estadístico de frecuencias que al procesamiento lingüístico. En los sistemas automáticos, las posibilidades de establecer relaciones de equivalencia y de jerarquía entre las categorías son bastante limitadas; el resultado suele ser una taxonomía plana, más próxima a un sistema de agrupación de recursos ("clustering") que de clasificación propiamente dicho. Un ejemplo de este tipo de soluciones es el módulo Automatic Taxonomy Generation de IDOL Server (http://www.autonomy.com/content/Products/IDOL).

Las soluciones completamente automáticas no han ofrecido, hasta el momento actual, resultados satisfactorios en lo que respecta a la construcción de taxonomías. En consecuencia, se están desarrollando alternativas semiautomáticas que, como Ultraseek Topic Advisor (http://www.verity.com/products/ultraseek/index.html) asiste con el proceso de creación y mantenimiento de la taxonomía a la vez que proporciona una interfaz para la revisión y aprobación de categorías. Dichos sistemas incluyen un algoritmo de base estadística que analiza un corpus de recursos y sugiere términos y relaciones entre términos al administrador del sistema para que éste los acepte o deniegue. Todo ello en un entorno amigable de trabajo.

3. Categorización de recursos

La categorización puede ser definida como el proceso de representación del contenido, contexto y/o estructura de recursos de información mediante la asignación de términos procedentes de un lenguaje documental -categorización por asignación- o mediante la extracción de términos de los propios recursos -categorización por extracción-.

El modelo de categorización más eficaz que existe en la actualidad es el que se basa en los metadatos. Siguiendo a Méndez y Senso (2004), podemos definir los metadatos como:

″ toda aquella información descriptiva sobre el contexto, calidad, condición o características de un recurso, dato u objeto que tiene la finalidad de facilitar su recuperación, autentificación, evaluación, preservación y/o interoperabilidad ″.

Existen diferentes modelos de metadatos. Los elementos que permiten establecer diferencias entre estos modelos son, básicamente, dos:

- Qué aspectos de los recursos que representan (los elementos).

- Cómo se representan dichos elementos (la sintaxis).

Por ejemplo, Dublin Core, uno de los modelos más utilizados para la descripción de todos los tipos de recursos de información, incluye, en su formato más sencillo ("nivel simple"), quince elementos [5] : Título, Autor o Creador, Claves, Descripción, Editor, Otros colaboradores, Fecha, Tipo de recurso, Formato, Identificador del recurso, Fuente, Lengua, Relación, Cobertura y Derechos. La sintaxis de cada elemento suele incorporar tres componentes:

- La identificación del elemento. Por ejemplo, en Dublin Core, el elemento Palabras clave se identifica mediante la metaetiqueta DC.Subject.

- Un o más calificadores que especifican algún atributo específico del elemento. Por ejemplo, un calificador de

la metaetiqueta DC.Subject puede ser SCHEME, que identifica el nombre del vocabulario controlado aplicado para la categorización del elemento.

- El valor o valores del elemento asignados al recurso que se describe. Por ejemplo, los términos extraídos del lenguaje controlado utilizado para la categorización del elemento.

En una página web codificada mediante el metalenguaje HTML, la sintaxis del elemento Claves presentaría el siguiente aspecto:

<META NAME="DC.Subject" SCHEME="TAGS" CONTENT="Herencia cultural; Acontecimientos culturales; Exposiciones; Gestión de documentación administrativa; Internet; Archivos; Gestión de la información">

En el modelo de categorización basado en metadatos, la taxonomía constituye un tipo de vocabulario controlado muy útil para la extracción de los valores -los términos- que se asignarán a los elementos que describen los recursos de información. Tal y como hemos indicado anteriormente, la aplicación de taxonomías no tiene que limitarse a los elementos que expresan el contenido de los recursos, y, más exactamente, a la materia, tema o disciplina. Los elementos relativos al contexto y a la estructura de los recursos también pueden ser expresados mediante categorías extraídas de una taxonomía.

La utilización de taxonomías en la categorización de recursos de información ofrece los puntos fuertes generales de los lenguajes controlados, como son: el tratamiento de los aspectos semánticos y sintácticos del lenguaje; la representación de conceptos implícitos; la creación de una visión global de los dominios que son objeto de representación; la exhaustividad en la indización; y la solución de los problemas que conllevan los contextos multilingües. Desde el punto de vista de la gestión de sitios web, la utilización de taxonomías en la categorización de los recursos ofrece dos importantes beneficios adicionales:

- Por un lado, rentabiliza los esfuerzos de construcción y mantenimiento de la taxonomía y de categorización de recursos; ya que una misma herramienta puede ser reutilizada en el desarrollo de diferentes aplicaciones de búsqueda, navegación, personalización, etc.

- Por otro lado, permite mantener la consistencia conceptual y designativa en la representación de los elementos de un mismo dominio, lo cual crea en los usuarios una imagen de consistencia en el conjunto del sitio web, y en la entidad que lo crea y lo mantiene.

El modelo de categorización aplicado por una organización determinada debe dar respuesta a cuatro cuestiones fundamentales: ¿qué recursos de información serán categorizados?, ¿con qué finalidad?, ¿quién los categorizará?, ¿cómo lo hará?

La dos últimas cuestiones están profundamente relacionadas con el grado de automatización aplicado a la asignación de valores a los metadatos. Desde este punto de vista, los sistemas de categorización se pueden concebir como un *continuum* , en uno de cuyos extremos se encuentran los sistemas totalmente manuales (mejor dicho, intelectuales) y, en el otro, los sistemas completamente automáticos.

En el primer caso, un experto analiza el contenido, el contexto, y/o la estructura de un recurso y le asigna las categorías oportunas a partir de un lenguaje controlado (categorización por asignación) o a partir del texto del propio recurso (categorización por extracción). La categorización intelectual ofrece, como puntos fuertes, un alto nivel de exactitud en la descripción de los recursos y la capacidad de incorporar el significado contextual en la descripción. Además, facilita la categorización de documentos no textuales (imágenes, aplicaciones, etc.); los puntos débiles son la limitada escalabilidad, el elevado coste en recursos humanos, y la falta de consistencia y exhaustividad.

Bennett (2002) presenta los siguientes datos relativos a los costes de la categorización manual de recursos en sitios web:

Yahoo!

- Dispone de una plantilla de 200 personas (aproximadamente) para la categorización de recursos.

- Utiliza una jerarquía de unas 500.000 categorías.

MEDLINE (National Library of Medicine)

- Invierte unos 2 millones de dólares al año en la indización manual de artículos de revista.

- Utiliza *MEdical Subject Headings* (18.000 categorías).

- Mayo Clinic

- Invierte 1's4 millones de dólares anuales para la codificación de acontecimientos médicos.

- Utiliza la International Classification of Diseases (ICD).

- US Census Bureau decennial census

- Debería invertir 15 millones de dólares para elaborar las respuestas de forma completamente manual.

- 232 categorías relativas a la industria y 504 categorías relativas a ocupación laboral.

La categorización automática se fundamenta en algoritmos que analizan estadísticamente las secuencias de palabras de los documentos, identifican patrones de comportamiento de las palabras a partir de variables como la colocación, orden, proximidad, frecuencia, etc., y agrupan los documentos que presentan similitud en dicho comportamiento. El resultado son agrupaciones ("clusters") de recursos que muestran patrones de comportamiento similares, etiquetadas mediante la secuencia de palabras extraídas de los propios recursos que mejor representan la similitud.

Un sistema de agrupación ha de ser capaz de realizar las siguientes tareas: analizar estadísticamente las secuencias de palabras de un recurso; computar el valor que representa numéricamente el contenido del documento; y comparar los valores de dos (sub)documentos y determinar su grado de similitud.

En el momento actual, los algoritmos diseñados para el análisis de frecuencias, utilizan algunos de los siguientes métodos de análisis, o una combinación de varios: métodos probabilísticos (método bayesiano, método de Rocchio...); métodos vectoriales (método K-Nearest Neighbor, Support Vector Machines...); y árboles y listas de decisión.

Como ejemplos de sistemas de categorización automática, pueden citarse el módulo Automatic Categorization de IDOL Server (http://www.autonomy.com/content/Products/IDOL), que se basa en el método probabilístico bayesiano, y Lotus Discovery Server (http://www.lotus.com), que se basa en el método vectorial [6] .

Los puntos fuertes de la categorización automática son la eficacia y rapidez de procesamiento, el alto nivel de escalabilidad y el alto nivel de consistencia; su gran punto débil es el bajo nivel de exactitud que suele ofrecer, lo que motiva que a menudo estos sistemas sean utilizados como base para la toma de decisiones por parte de categorizadores humanos.

Los sistemas de categorización semiautomática o híbrida combinan la inteligencia humana, que puede identificar los diferentes niveles de significado existentes en los documentos, y la eficiencia de los automatismos. Se pueden identificar cuatro familias de sistemas semiautomáticos de categorización.

- Sistemas que analizan estadísticamente los recursos y presentan a los expertos humanos términos recomendados de categorización para que éstos los revisen y aprueben. Un ejemplo de este tipo de sistemas es Ultraseek Advanced Classifier (http://www.verity.com/products/ultraseek/index.html).

- Sistemas de categorización basada en reglas de búsqueda. Permite vincular a cada una de las categorías de una taxonomía una ecuación de búsqueda diseñada por especialistas mediante opciones avanzadas (regla de búsqueda). Mediante un algoritmo, el sistema analiza los documentos y determina cuál o cuáles son las ecuaciones con las que manifiesta mayor coincidencia. A continuación, asigna el documento a la categoría o categorías que tienen vinculadas dichas reglas de búsqueda. Son ejemplos de este tipo de sistemas K2 Enterprise [7] (http://www.verity.com/products/k2_enterprise/index.html) y Ultraseek Content Classification Engine (http://www.verity.com/products/ultraseek/cce.html), ambos de Verity.

- Sistemas de categorización basada en conjuntos de documentos de entrenamiento o ejemplares. Permite vincular a cada una de las categorías de una taxonomía un número limitado de documentos seleccionados por especialistas que son considerados los más relevantes. Mediante un algoritmo, el sistema analiza los nuevos documentos que deben ser categorizados y determina a qué documentos ejemplares se aproxima más. A continuación, asigna el documento a la categoría o categorías de los más relevantes. Un ejemplo de este tipo de sistemas es Mohomine Classifier (http://www.kofax.com/products/mohomine/classifier.asp), de Mohomine.

- Sistemas de categorización basada en el análisis lingüístico. Un ejemplo de este tipo de sistemas es Smart Discovery [8] de InXight.

Los puntos fuertes de los sistemas de categorización semiautomáticos son un buen equilibrio entre eficiencia y exactitud; el hecho de que el proceso esté guiado por el razonamiento humano; y la capacidad de acumular y generar aprendizaje. Entre los puntos débiles, cabe destacar la exigencia de conocimientos, habilidades y esfuerzos de gestión y mantenimiento.

En una encuesta realizada por Delphi Research [9] , los directivos de 300 grandes empresas de todo el mundo (el 60%, norteamericanas) dieron las siguientes respuestas a la pregunta sobre el tipo de estrategia de implementación de la taxonomía: el 36%, híbrida; el 26%, automática; el 23%, manual; el resto, o bien otras opciones o no dieron respuesta alguna.

4. Aplicación de la taxonomía en el desarrollo de sistemas de búsqueda de información

Como ya se ha indicado anteriormente, la diferenciación de los procesos de creación de la taxonomía, de categorización de recursos mediante las categorías de la taxonomía y de aplicación de la taxonomía ofrece múltiples beneficios. El objetivo de la construcción de ésta es representar una realidad (un área de conocimiento, el ámbito de actividad de una organización, etc.) de la forma más adecuada a los propósitos e intereses de la entidad que debe explotar dicha representación. Además, debe constituir expresión de la imagen e intereses corporativos de la propia entidad.

Las aplicaciones de la taxonomía en el contexto de los sitios web pueden ser diversas; si nos centramos al ámbito de la arquitectura de la información, una misma taxonomía puede constituir una herramienta básica o auxiliar para los diferentes sistemas de navegación, de organización y búsqueda de contenidos, de etiquetado, y de personalización. La reutilización de una misma taxonomía para diferentes herramientas de arquitectura de información ofrece diferentes beneficios:

- En primer lugar, permite la rentabilización del esfuerzo inicial de creación de la taxonomía y de los esfuerzos subsiguientes de mantenimiento.

- En segundo lugar, facilita la gestión de las funcionalidades que aplica la taxonomía: una modificación en las categorías o en las relaciones entre categorías de la taxonomía puede trasladarse uniforme y consistentemente a todas las funcionalidades.

- En tercer lugar, mejora el uso del sitio web en su conjunto ya que reduce considerablemente las exigencia de carga cognitiva, de memoria y de aprendizaje.

- En cuarto lugar, facilita la interacción con el sitio web y la creación de una imagen consistente de la organización que crea y aplica la taxonomía.

Existen diferentes opciones de presentación de la taxonomía.

- Presentación íntegra de la taxonomía, con todas sus categorías y las relaciones que las interconectan (relación de equivalencia, modelo estructural jerárquico o facetado, etc.).

- Presentación parcial de la taxonomía original, para destacar contenidos a partir de criterios temporales o de uso.

- Reducción de la taxonomía a la relación de equivalencia, de forma que la taxonomía adopta la forma de anillo de sinónimos.

- Reducción de la taxonomía a la relación jerárquica, para su utilización como sistema de exploración de categorías. En este caso, suele comportar la reducción de los niveles de amplitud y de profundidad para ajustar la taxonomía a las recomendaciones derivadas de las limitaciones de capacidad cognitiva, visual y de memoria del usuario estándar.

- Presentaciones alternativas, como pueden ser la ordenación alfabética de las categorías, o las presentaciones arbórea, gráfica y metafórica.

La selección de una opción depende de diversos factores; la funcionalidad para la que se aplica, los usuarios a los que se dirige, etc. Generalmente, la combinación de diferentes presentaciones en una misma funcionalidad ofrece buenos resultados.

Una de las funcionalidades de los sitios web en los que la taxonomía juega un papel protagonista es la búsqueda de información. Los sistemas que permiten buscar contenidos en el entorno web pueden clasificarse en tres grandes tipos: de exploración ("browsing"), de recuperación ("searching") y de filtraje ("filtering").

Los sistemas de búsqueda por exploración ofrecen a los usuarios una estructura organizada de categorías donde se incorporan los recursos de información, y un mecanismo de navegación por dichas categorías para localizar los recursos relevantes para sus necesidades de información. Estos sistemas de exploración son especialmente convenientes para situaciones de búsqueda en que los usuarios no pueden concretar excesivamente la necesidad de información (búsqueda exploratoria). El mecanismo de navegación puede ser:

- La estructura jerárquica o facetada original de la taxonomía, completa o reducida.

- Una de las presentaciones alternativas que hemos indicado anteriormente: alfabética, arbórea, gráfica o metafórica.

- La combinación de dos o más de estas presentaciones de forma que el usuario pueda seleccionar la que más convenga a las condiciones de su necesidad de información.

Los sistemas de recuperación de información ofrecen a los usuarios la posibilidad de crear una ecuación de búsqueda a partir de una palabra o una combinación de palabras. Estos sistemas de exploración son especialmente convenientes para situaciones de búsqueda en que los usuarios pueden concretar con suficiente detalle la necesidad de información (búsqueda de ítem conocido). La taxonomía se incorpora al sistema de recuperación para auxiliar al usuario en la identificación de términos relevantes para la creación de la ecuación de búsqueda, y también para mejorar los procesos de presentación de resultados y reformulación de la consulta. Los sistemas de exploración y de recuperación suponen la interacción a tiempo real entre el usuario y el mecanismo de búsqueda.

La tercera modalidad, los sistemas de filtraje, ofrece la posibilidad al usuario de crear y declarar una necesidad de información (perfil de usuario), y recibir una respuesta automática cuando se cumple un plazo determinado o cuando el sistema identifica recursos relevantes para dicha necesidad. En este caso, la taxonomía permite al usuario seleccionar términos relevantes para la concreción de su perfil.

5. Bibliografía

Bennett, Paul. (2002). Introduction to text categorization. Consultado: 1-mar-2005, http://www.softlab.ece.ntua.gr/facilities/public/AD/Text% 20Categorization/Introduction%20to%20Text%20Categorization .ppt#256 , 1, Introduction to Text Categorization.

Diccionario de la lengua española (2001). Consultado: 22-mar-2005, http://buscon.rae.es/diccionario/drae.htm

Fast, Karl; Leise, Fred; Steckel, Mike (2003). "Controlled vocabularies: a glosso-thesaurus". En: Boxes & arrows, October 27, 2003. http://www.boxesandarrows.com/archives/controlled_vocabularies _a_glossothesaurus.php

Gilchrist, Alan; Kibby, Peter; Mahon, Barry. (2000). *Taxonomies for business: access and connectivity in a wired world*. London: TFPL. ISBN: 1-870-889-83-5.

Grove, Andrew. "Taxonomy". (2003). En: *Encyclopedia of library and information science*. 2nd ed., rev and enlarg. New York [etc.]: Marcel Dekker, p. 2770-2777.

IDOL Server. (2005). Consultado: 13-mar-2005, http://www.autonomy.com/content/Products/IDOL

Information intelligence: content classification and the enterprise taxonomy practice (2004). Consultado: 25-ene-2005, http://www.delphigroup.com/research/whitepapers/20040601 -taxonomy-WP.pdf

K2 Enterprise. (2005). Consultado: 13-mar-2005, http://www.verity.com/products/k2_enterprise/index.html

Lotus Discovery Server. (2004). Consultado: 1-sep-2004, http://www.lotus.com

Mathes, Adam. (2004). Folksonomies: cooperative classification and communication through shared metadata. Consultado: 26-gen-2005, http://www.adammathes.com/academic/computer-mediated -communication/folksonomies.html

Méndez, Eva; Senso, José A. (2004). Introducción a los metadatos. Consultado: 14-ene-2004, http://www.sedic.es/autoformacion/metadatos/introduccion.htm

Metainformación: Dublin Core. (2003). Consultado: 13-mar-2005, http://www.rediris.es/metadata

Mohomine Classifier. (2005). Consultado: 13-mar-2005, http://www.kofax.com/products/mohomine/classifier.asp

Multites 2005. (2005). Consultado: 13-mar-2005, http://www.multites.com

National Information Standards Organization. (2005). ANSI/NISO Z39.19-2003: guidelines for the construction, format, and management of monolingual thesauri. Consultado: 9-mar-2005,

http://www.niso.org/standards/standard_gather.cfm?pdflink =http://www.niso.org/standards/resources/Z39-19.pdf&std_id=518 [Consulta: 9-3-2005].

Ruiz, Miguel E.; Srinivasan, Padmini. "Combining machine learning and hierarchical indexing structures for text categorization". En: ASIS/SIGCR Workshop on Classification Research (10è: Washington: 1999). Advances in classification research: proceedings of the ASIS SIG/CR Classification Research Workshop, v. 10 (1999), p. 107-124.

Smart Discovery. (2005). Consultado: 13-mar-2005, http://www.inxight.com/products/smartdiscovery

"TAG Conference Call, may 19, 2003" (2003). En: National Information Standards Organization. (2004). Developing the next generation of standards for controlled vocabularies and thesauri. Consultado: 23-abr-2004. http://www.niso.org/committees/MTinfo.html

"TAG Conference Call, June 30, 2003" (2003). En: National Information Standards Organization. (2004). Developing the next generation of standards for controlled vocabularies and thesauri. Consultado: 23-abr-2004. http://www.niso.org/committees/MTinfo.html

"TAG Notes November 1, 2004" (2004). En: National Information Standards Organization. (2004). Developing the next generation of standards for controlled vocabularies and thesauri. Consultado: 23-abr-2004. http://www.niso.org/committees/MTinfo.html

Taxonomy strategies. Consultado: 25-ene.-2005, http://www.taxonomystrategies.com/index.htm

Taxonomy warehouse. Consultado: 22-feb.-2005, http://www.taxonomywarehouse.com

Term Tree. (2005). Consultado: 13-mar-2005, http://www.termtree.com.au

Ultraseek Advanced Classifier. (2005). Consultado: 22-feb.-2005, http://www.verity.com/products/ultraseek/index.html

Ultraseek Content Classification Engine (CCE). (2005). Consultado: 13-mar-2005, http://www.verity.com/products/ultraseek/cce.html

Ultraseek Topic Advisor. (2005). Consultado: 22-feb.-2005, http://www.verity.com/products/ultraseek/index.html

Webopedia. Consultado: 28-ene.-2005, http://www.pcwebopedia.com/TERM/t/taxonomy.htm

Miquel Centelles forma parte del Grupo DigiDoc del Instituto Universitario de Lingüística Aplicada de la Universidad Pompeu Fabra . Este artículo presenta una parte de los resultados del proyecto HUM2004-03162/FILO del Plan Nacional I+D+I del Ministerio de Educación y Ciencia (España)

6. Notas

[1] De acuerdo con las "TAG Notes November 1, 2004" (2004), el borrador final había de estar preparado para enero del 2005. [volver]

[2] Se puede obtener una copia de las referencias, enviando un mensaje de correo electrónico al autor de este artículo (miquel.centelles@ub.edu). Es necesario especificar el motivo de la solicitud. [volver]

[3] En estos casos, el término taxonomía se acompaña del calificativo "corporativa". [volver]

[4] Un ejemplo de esta opción es SemioTaxonomy de la empresa Entrieva. Más información en: http://www.entrieva.com/entrieva/products/scts.asp?Hdr=scts [Consultado: 13-mar-2005]- [volver]

[5] Información extraída del sitio web Metainformación: Dublin Core (2003), mantenido por RedIRIS. [volver]

[6] Según el informe Information intelligence: content classification and the enterprise taxonomy practice (2004, p. 38), Autonomy ocupa una cuota de mercado del 14% y Lotus Discovery Server del 7%. [volver]

[7] Según el informe Information intelligence: content classification and the enterprise taxonomy practice (2004, p. 38), K2 ocupa una cuota de mercado del 15%. [volver]

[8] Según el informe Information intelligence: content classification and the enterprise taxonomy practice (2004, p. 38), Smart Discovery ocupa una cuota de mercado del 4%. [volver]

[9] Information intelligence: content classification and the enterprise taxonomy practice (2004, p. 26). [volver]