TESAUROS Y ONTOLOGÍAS EN SISTEMAS DE INFORMACIÓN DOCUMENTAL

Lluís Codina y Rafael Pedraza-Jiménez

Lluís Codina es profesor titular en el *Departamento de Comunicación* de la *Universidad Pompeu Fabra* (*UPF*). Imparte docencia en los *Estudios de Periodismo y de Comunicación Audiovisual* de la *UPF*. Participa en masters oficiales y en programas de doctorado del *Departamento de Comunicación* y del *Instituto de Educación Continuada* (*IDEC/UPF*). Coordina el *Grupo de Investigación en Documentación Digital y Comunicación Interactiva*. Es fundador y codirector del primer máster que se imparte íntegramente en la Web sobre documentación digital.

Departamento de Comunicación. Universidad Pompeu Fabra
Roc Boronat, 138. 08018 Barcelona
Tel.: +34-935 421 311
lluis.codina@upf.edu
http://www.lluiscodina.com

Rafael Pedraza-Jiménez es doctor en documentación por la *Universidad de Barcelona* (*UB*) y profesor del *Depto. de Comunicación* de la *Universidad Pompeu Fabra*, donde imparte docencia de documentación en los estudios de *Periodismo*, *Comunicación Audiovisual*, y *Publicidad y Relaciones Públicas*. Asimismo, colabora como profesor en diferentes masters oficiales, de la *UB*, o el *Máster Calsi* de la *Univ. Politécnica de Valencia*, así como en el *Máster de Gestión de Contenidos Digitales IDEC/UPF*. Investiga en Web semántica y recuperación de información.

Departamento de Comunicación. Universidad Pompeu Fabra
Roc Boronat, 138. 08018 Barcelona
Tel.: +34-935 422 437
rafael.pedraza@upf.edu

Resumen

Desde la (re)aparición de las ontologías a finales de los noventa, existe un debate sobre cuál es la relación entre tesauros y ontologías. Este artículo pretende mostrar una síntesis de este debate centrándose en su uso en los sistemas de información, el mejor terreno para comparar ambas tecnologías utilizando a la vez el paraguas conceptual de la semántica documental.

Palabras clave

Tesauros, Ontologías, Sistemas de información, Semántica documental

Title: Ontologies and thesauri in information systems

Abstract

Since the (re)appearance of ontologies in the late nineties, there is a debate about the relationship between thesauri and ontologies. This article presents a synthesis of this debate, focusing on the use of these tools in the information systems area, which is the best one to compare both technologies, utilizing the concept of document description semantics.

Keywords

Thesauri, Ontologies, Information systems, Document description semantics.

Codina, Lluís; **Pedraza-Jiménez, Rafael**. "Tesauros y ontologías en sistemas de información documental". *El profesional de la información*, 2011, septiembre-octubre, v. 20, n. 5, pp. 555-563.
http://dx.doi.org/10.3145/epi.2011.sep.10

1. Introducción

La semántica documental forma parte del ADN de las ciencias de la documentación como pocas otras cosas. En este trabajo pretendemos ocuparnos de uno de sus instrumentos más importantes: los tesauros, a la luz de una comparación con el instrumento a su vez más significativo de la web semántica: las ontologías.

El motivo de tal comparación es que las ontologías pueden considerarse también bajo la óptica de la semántica documental, ya que constituyen una nueva forma de represen-

Artículo recibido el 15-08-11
Aceptación definitiva: 21-08-11

La semántica documental y la web semántica

Podemos definir la semántica documental como el campo de estudios y de aplicaciones profesionales vinculado con la representación de documentos, tanto con la representación de su contenido (información) como con su identificación y descripción en tanto que objetos (soporte).

Hasta ahora la semántica documental disponía principalmente de dos tipos de instrumentos: los lenguajes documentales (para describir el contenido), y los esquemas de metadatos (para identificar el soporte o el objeto en sí). Por ejemplo, una fotografía en un banco de imágenes se representa mediante dos conjuntos de datos: un grupo de palabras clave o descriptores extraídos de una taxonomía o de un tesauro más un grupo de metadatos que incluye componentes como el título, el autor y lo derechos de la imagen.

Por tanto, hasta ahora, se podía representar este campo con la siguiente ecuación:

Semántica documental = Lenguajes documentales (LD) + Esquemas de metadatos (EM)

A partir de la creciente importancia real de la web semántica y del protagonismo de las ontologías en su seno, ahora sería más adecuado representarlo de esta forma:

Semántica documental = Lenguajes de representación del conocimiento (LRC) + Esquemas de metadatos (EM)

Formarían parte de los LRC tanto los tesauros como las ontologías, aspecto que ya ha sido puesto de relieve por diversos autores que consideran que existe un *continuum* con las clasificaciones y taxonomías en un extremo y las ontologías en el otro, y los tesauros en algún lugar intermedio del mismo (**McGuinness, Deborah L.** "Ontologies come of age". Fensel, D. *et al* (ed.). *Spinning the semantic web*. Cambridge: The MIT Press, 2005).

A su vez, formarían parte de los esquemas de metadatos tanto las habituales y familiares reglas de catalogación (p. e. las AACR2) como los sistemas relativamente nuevos (p. e. Dublin Core).

tación de la información que combina a la vez las características de un lenguaje documental (como los tesauros o las taxonomías) y de un sistema de metadatos (como el sistema Dublin Core o las AACR2).

Diversos autores se han ocupado anteriormente (ver referencias) de presentar otras comparaciones del tipo tesauros versus ontologías (incluso existe una entrada específica de la *Wikipedia* consagrada a ello). Sin embargo, esos análisis suelen centrarse en analizar las ventajas *teóricas* de uno y otro, pero sin contrastarlas con ningún contexto común.

Entendemos que tal tipo de análisis era necesario y oportuno en su momento, pero más de una década después del nacimiento oficioso de la web semántica (como proyecto) parece oportuno recuperar ese análisis ahora sobre una base más real y, dados los intereses de la biblioteconomía-documentación, más relacionado con la semántica documental.

Además, aunque no todos, entendemos que algunos de los análisis anteriores, tal vez de manera inevitable, presentaban las ontologías de una forma un tanto acrítica y aislada del mundo real, al dar por válida la visión de los organismos y desarrolladores involucrados oficialmente en el proyecto de la web semántica.

Aquí hay que señalar enseguida una primera constatación: mientras todo lo referido a tesauros se mueve en el terreno de la más estricta realidad cotidiana, casi todo lo referido a ontologías se mueve todavía en el terreno de una operación de prospectiva no exenta de especulación por mucho que pueda tener bases más o menos sólidas.

En lo que sigue proponemos una comparación entre ambos instrumentos, tesauros y ontologías, en el contexto de la semántica documental y siempre desde el punto de vista de sus posibilidades de utilización en sistemas de información. Pero antes puede ser útil recordar el debate sobre el uso de lenguaje libre o de lenguaje controlado en la indización de documentos.

2. Lenguaje libre versus lenguaje controlado

Antes de seguir deberíamos revisar el papel de los lenguajes documentales (LDs a partir de ahora) en la organización de la información. Como es sabido, entre otros sistemas propios de la semántica documental, los tesauros ayudan a caracterizar las propiedades de los documentos por lo que se refiere a su contenido, asociando a los mismos grupos de descriptores elegidos de dichos tesauros.

Por tanto, el papel de los lenguajes documentales y, por ello, de los tesauros, también se puede expresar diciendo que consiste en proporcionar un lenguaje controlado que no depende de la lengua natural del documento sino de una serie de términos normalizados (descriptores).

En lugar de términos normalizados (descriptores) ¿por qué no utilizar la lengua natural del documento como hacen, por ejemplo, los motores de búsqueda? Hay varias razones. Las más importantes podrían ser las siguientes:

– En primer lugar, porque para aplicar una indización en lenguaje natural se necesitan documentos de texto completo, y en algunos sistemas documentales se carece de los mismos. Es el caso de las bases de datos académicas de tipo referencial. En esta clase de sistemas únicamente se puede trabajar con los datos de la referencia bibliográfica y a lo sumo un pequeño resumen. En estos casos es necesario un lenguaje controlado como el que aportan los tesauros.
– En segundo lugar, porque algunos sistemas de información contienen documentos audiovisuales (y no textuales) como fotografías y vídeo. En tales casos el texto, cuando lo hay, es un mero complemento que contiene aún menos información que en el caso de las bases de datos referenciales.
– Por último, en determinados entornos, aunque haya documentos de texto completo se ha observado que el sistema incrementa mucho sus prestaciones de cara a los usuarios si, además de la indización por texto completo,

se puede utilizar un lenguaje controlado que ayude a los usuarios a explorar el fondo documental.

Adicionalmente, podemos señalar que una razón por la cual los motores de búsqueda no parecen necesitar utilizar tesauros (u otras formas de semántica documental, como ontologías) es porque disponen de un arma formidable que no se da en otros contextos (p. e., en bancos de imágenes o en bases de datos de artículos de revista, etc.) a saber, el análisis de enlaces.

Basta recordar aquí que, hasta que buscadores como *Google* no empezaron a aplicar el análisis de enlaces como forma para determinar la relevancia de las páginas web, su eficacia era tan limitada que muchos usuarios preferían utilizar otros sistemas, por ejemplo, los directorios o las bases de datos de recursos de internet (aún hoy funcionan varios de ellos, como *Intute*), *http://www.intute.ac.uk*

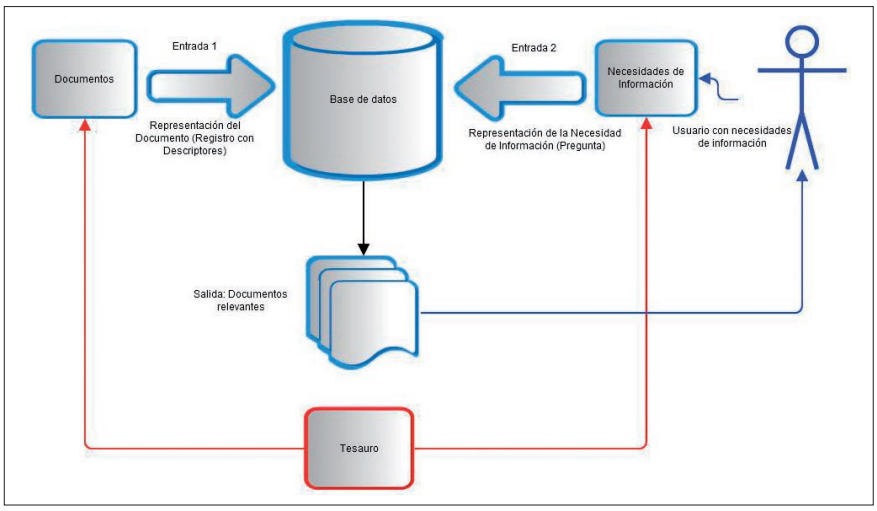

Figura 1. El papel de los lenguajes documentales (tesauros) en un sistema de información documental

Por tanto, en algunos casos (bases de datos referenciales, bases de datos de objetos no textuales, fondos documentales sin hiperenlaces, etc.) la indización basada en el texto libre o lenguaje natural o bien es insuficiente o bien es directamente imposible. Es en esta clase de sistemas donde son necesarios LDs como los tesauros.

Ahora bien, hay al menos cuatro términos que se manejan de forma muy problemática como si fueran sinónimos, y son los siguientes: clasificaciones, taxonomías, tesauros, ontologías. Entendemos que un uso consensuable de los tres primeros es el mostrado en la tabla 1 (nos guiamos principalmente por la norma *NISO Z39.19-2005*).

3. Tesauros

Por otro lado, los tesauros sirven también para representar las necesidades de información de los usuarios, o sea las preguntas que formulan al sistema (los usuarios indizan las preguntas). Por lo tanto, los tesauros son intermediarios en un proceso de información que ayuda al usuario a representar sus necesidades de información mediante el mismo sistema de descriptores que, previamente, sirvió para indizar o representar el contenido de los documentos.

Vemos, pues, que los tesauros se pueden usar para caracterizar las dos entradas de todo sistema de información documental: documentos y preguntas. Veamos a continuación dos ejemplos, el primero de ellos con la base de datos *ISOC* del *CSIC* y el segundo con la base de datos *ERIC* del *Department of Education* de Estados Unidos.

Base de datos *ISOC*

Está producida por el *Centro de Ciencias Humanas y Sociales* del *Consejo Superior de Investigaciones Científicas* (*CSIC*). El *CSIC* es una agencia estatal del *Ministerio de Ciencia e Innovación*, y es la mayor institución pública dedicada a la investigación en España y la tercera de Europa. Una parte de sus bases de datos son de acceso libre, como la que nos ocupa aquí, la sección de la base de datos *ISOC* dedicada a temas de biblioteconomía y documentación.
http://www.csic.es
http://www.cchs.csic.es
http://bddoc.csic.es:8080/isoc.html

Una vez en esta base de datos, se puede hacer una búsqueda simple utilizando el formulario que aparece como primera opción tal como podemos ver en la figura 2.

Si se realiza una consulta, por ejemplo utilizando el término Ontologías y se selecciona uno de los registros, se obtendrá un resultado similar al que se reproduce en la figura 3.

En la zona vertical (color azul) aparecen destacados los nombres de los campos, cada uno de ellos una clase de metadatos. Lo que se ha destacado en horizontal (color rojo)

Clasificación	En general es una estructura de organización jerárquica formada por clases y subclases. En el ámbito de la documentación es un conjunto de términos organizados en forma jerárquica, sin necesidad de incluir relaciones explícitas o formales (como términos relacionados, término genérico, etc.). Puede incluir relaciones de equivalencia (término preferido-no preferido).
Taxonomía	En rigor, una taxonomía es una clasificación. Por una deriva semántica que posiblemente obedece a influencias de prestigio del término, ya que procede de la biología, de donde lo tomó a su vez la informática, es utilizado como sinónimo de lenguaje controlado en general, sobre todo en determinadas áreas como en arquitectura de la información y también en sistemas de información no bibliotecarios (p. e., en sistemas de información corporativos).
Tesauro	Conjunto de términos preferidos (descriptores) y no preferidos (sinónimos) utilizados para representar un campo del conocimiento y generalmente para representar el contenido de los documentos de una base de datos o sistema de información. Debe contener al menos las siguientes relaciones entre términos: equivalencia, jerarquía y asociación. Existen normas nacionales (*UNE*) e internacionales (*NISO*) que definen sin ambigüedad la estructura de un tesauro, y para merecer ese nombre un lenguaje documental debería ajustarse a ellas.

Tabla 1. Definiciones de los conceptos clasificación, taxonomía y tesauro (elaboradas a partir de la norma *ANSI/NISO 2005*)

Figura 2. Formulario de búsqueda simple de la base de datos *ISOC/CSIC*

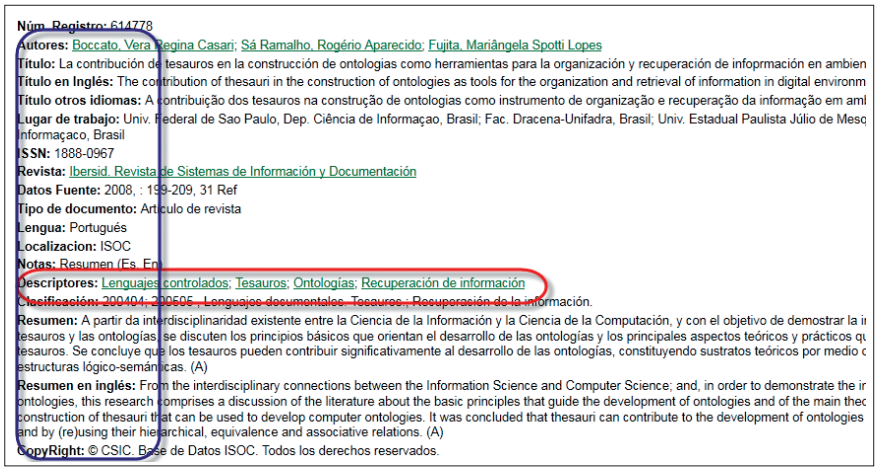

Figura 3. Ejemplo de registro de la base de datos *ISOC/CSIC*

son los descriptores con los cuales se ha indizado este documento. En concreto vemos los siguientes: Lenguajes controlados, Tesauros, Ontologías, Recuperación de Información. Esto significa que, si el usuario no había pensado en alguno de los descriptores, p. e, no había pensado en Lenguajes controlados, Tesauros, o en Recuperación de Información (recordemos que la pregunta era por Ontologías), si hace clic en alguno de estos descriptores encontrará todos los registros relacionados con ese concepto, incluso aunque el documento no lo contuviera (p. e., es posible que a un documento sobre Clasificaciones se le haya asignado el descriptor Lenguajes controlados, aunque en el documento en sí no se mencione ese término).

Con lo anterior hemos visto de forma fácil de qué modo un tesauro puede ser utilizado por los usuarios para mejorar sus opciones de recuperación de información. No obstante, hay algunas bases de datos que dan un paso más y permiten la consulta directa del tesauro como parte de su forma de hacer consultas. La base de datos del CSIC también permite el uso de su lenguaje controlado para lanzar consultas, pero de una forma no tan directa, a través de sus diversos índices. Para mostrar un modelo más claro de utilización del tesauro en consultas directas consultaremos *ERIC*.

Base de datos *ERIC*

ERIC (*Education Resources Information Center*) es una base de datos producida por el *Departamento* (ministerio) *de Educación* del gobierno de Estados Unidos. Probablemente es la base de datos más importante en temas de educación y tiene usuarios en todo el mundo.

Lo que nos interesa de este caso es que una de las opciones de búsqueda es mediante el tesauro, como podemos ver en la figura 4.

Una vez se hace clic en el enlace Thesaurus, el sistema permite hacer una doble operación. Por un lado se podrá consultar el tesauro en toda su extensión y con todas las relaciones de cada término (sinónimos, términos más generales, más específicos, relacionados). Por otro, se podrá además lanzar una búsqueda, incluso utilizando operadores booleanos junto con los descriptores del tesauro, como muestra la figura 5.

4. Ontologías

Las ontologías son una de las tecnologías más prometedoras para el futuro de los sistemas de información. El problema es que buena parte de las certezas que tenemos en el mundo de los tesauros como modelo de éxito en el campo de la semántica documental desaparecen en el caso de las ontologías; al menos cuando consideramos su aplicación a los sistemas de búsqueda y obtención de información.

A continuación se reflexionará sobre un tipo especial de problemas antes de pasar a considerar sus posibles formas de utilización en sistemas de información. Los problemas a los que nos referimos son, al menos, los tres siguientes:

– Imprecisión
– Insuficiencia
– Indeterminación

4.1. Imprecisión

La imprecisión del concepto ontología tiene al menos tres dimensiones: en primer lugar, en el uso real del término en la bibliografía especializada no existe todavía un consenso bien establecido, a semejanza del concepto de tesauro, de qué es una ontología. En segundo lugar, dado lo anterior, no es extraño que no exista aún un consenso sobre cómo elaborar una ontología, a diferencia una vez más de los tesauros. Por último, no existe un consenso amplio sobre qué componentes tiene una ontología.

Afortunadamente, el proyecto de la web semántica está consiguiendo asentar lo que podríamos denominar el sentido *canónico* del término ontología (**Noy**, 2001; **Pedraza**; **Codina**; **Rovira**, 2007). Si este asentamiento tiene éxito, es posible que en el futuro todas o parte de las tres dimensiones de imprecisión señaladas desaparezcan, pero actualmente no es así. De hecho, no es difícil encontrar publicaciones o interlocutores para los cuales, no ya un tesauro, sino una simple clasificación jerárquica es una ontología.

Dicho lo anterior, parecería una contradicción querer presentar ahora una única visión compresiva de qué puede

ser una ontología. La única forma de resolverlo (sin tener que presentar decenas de puntos de vista distintos y no acordar nada al final) es ceñirnos al concepto teórico –y pragmático a la vez– que se deriva de los lenguajes y recomendaciones del *W3 Consortium* en relación con la web semántica (al fin y al cabo ha sido gracias a este proyecto que el interés por las ontologías ha renacido, por así decirlo). Es por eso que hablamos del verdadero canon que las recomendaciones del *W3C* están estableciendo en este campo.

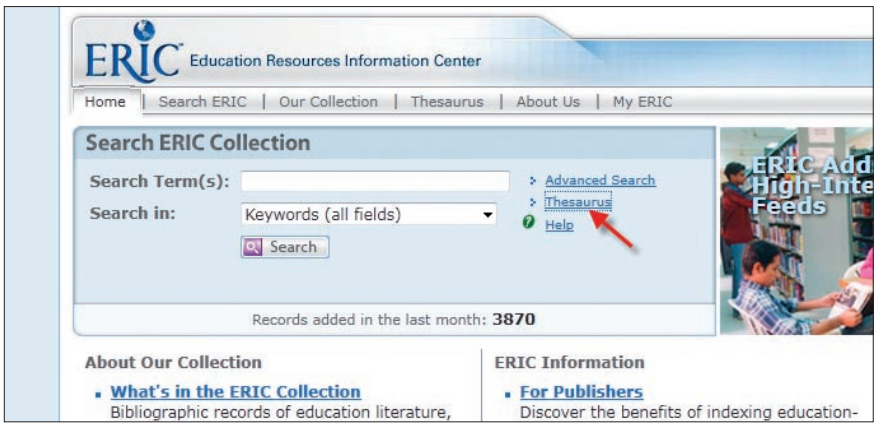

Figura 4. Página principal de la base de datos *ERIC* con acceso directo al tesauro (*Thesaurus*)

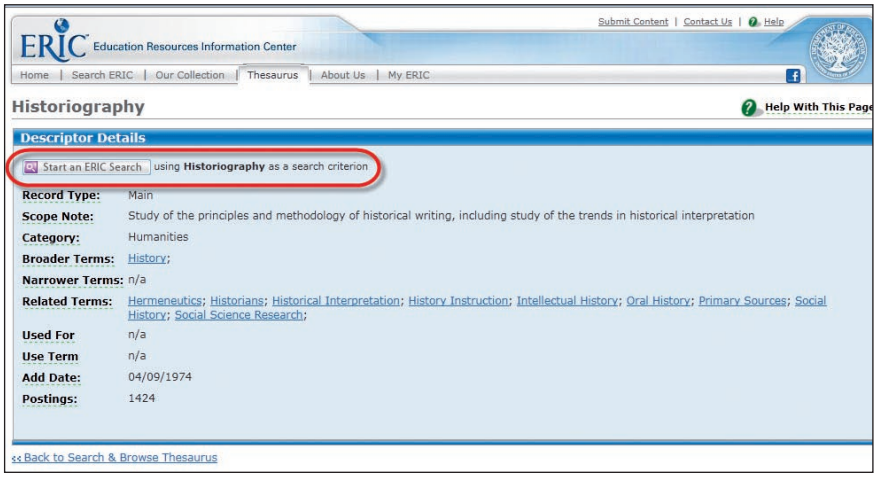

Figura 5. La presentación completa (ficha) de cada descriptor incluye navegación entre los términos y la posibilidad de lanzar consultas por alguno de ellos

La definición más breve, y a la vez probablemente la más citada, se debe a **Gruber** (1993), según la cual "una ontología es la especificación formal de un ámbito del conocimiento". Como es una definición tan compacta, convendrá discutir sus dos partes principales (empezando por el final):

– Un ámbito del conocimiento es cualquier entidad del mundo real o conceptual, simple o compleja, que se pueda concebir o conocer de una forma más o menos sofisticada. Al igual que los tesauros, las ontologías representan siempre algún tipo de conocimiento. En principio, cualquier tipo de conocimiento que se expresa en un tesauro puede expresarse en una ontología (lo contrario no sería cierto). Tal como existen tesauros para representar zonas geográficas (con regiones, países, ciudades y sus relaciones respectivas) pueden existir ontologías para representar esas mismas zonas (o todo el planeta). Tal como existen tesauros sobre economía, en tanto que ámbito del conocimiento, también puede haber ontologías sobre economía (o sobre cine, o sobre la mafia, o sobre medicina, o sobre los productos de una empresa, o sobre tipos de pizzas, etc.)
– Especificación formal: significa que una ontología debe quedar especificada siguiendo las exigencias de un formalismo bien determinado. Los tesauros también deben seguir un determinado formalismo (p. e. el que marca la norma *NISO*), pero en las ontologías la exigencia es muy superior, ya que, en primer lugar, las ontologías deben especificarse mediante un lenguaje informático con una fuerte base lógica-matemática (derivada de las llamadas lógicas descriptivas). Además, para que las ontologías sean algo más que una taxonomía codificada en un lenguaje complicado deben incluir los llamados axiomas. Éstos consisten en especificaciones, siempre en base lógica, de las propiedades y de las relaciones entre los componentes de la ontología. Por ejemplo, los ríos tienen afluentes, algunos ríos son navegables. O bien, dos personas que comparten los mismos progenitores, son hermanos; cada país sólo puede tener una capital, los países sin litoral no pueden tener puertos de mar, etc.

Retomando la definición de **Gruber**, si conseguimos especificar los componentes y sus relaciones de un ámbito del conocimiento siguiendo un formalismo estricto codificado en un lenguaje informático (no de programación, sino de descripción), entonces tenemos una ontología.

Por lo tanto, los componentes principales de una ontología son clases (p. e. vías fluviales, mamíferos), subclases (p. e. ríos, chimpancés) e individuos o instancias (p. e., el río Guadalquivir, la mona Chita), más sus relaciones (las propiedades se consideran un tipo de relación).

En segundo lugar, debe ser capaz de soportar pruebas lógicas. Es decir, una ontología no merece este nombre, al menos en el sentido canónico, si un programa de ordenador no puede realizar inferencias válidas sobre ella. Es decir, si en la ontología hemos definido que un individuo X pertenece a la clase B, la cual a su vez es miembro de la clase A, el ordenador debería poder inferir que todo miembro de B es un miembro de A, y que por tanto X es un A.

4.2. Insuficiencia

La insuficiencia, siempre en el terreno de los sistemas de información, se refiere a lo siguiente: aunque un tesauro es autosuficiente, una ontología no. Es decir, un tesauro impreso en papel tiene sentido. Por ejemplo, puede ser una guía para que el editor de una base de datos asigne descriptores a los registros de la misma, o para que un usuario obtenga ideas para su búsqueda.

En cambio, una ontología por sí sola en principio no tiene sentido si no hay un sistema complementario que permita realizar inferencias sobre la misma para, por ejemplo, facilitar búsquedas en lenguaje natural. Si "solamente" tenemos una ontología, a efectos de recuperación resuelve más bien poco (siempre en el contexto de los sistemas de información) aunque solamente sea porque no diferencia entre términos preferidos y no preferidos; tampoco establece relaciones asociativas ya que no tienen entrada en la lógica de base matemática de la ontología, no ofrece notas de alcance, etc.

Por el contrario, el ejemplo de una ontología útil, es decir, complementada con un sistema de inferencias debería soportar la siguiente búsqueda enunciada en lenguaje natural: próximas convocatorias abiertas para solicitar ayudas, becas o premios para producciones audiovisuales en países del Mediterráneo.

Para un ser humano es fácil entender que una convocatoria de ayudas a producciones de televisivas en España, entra en la categoría de país del Mediterráneo y de producción audiovisual. Para un ordenador (sin una ontología y un sistema de inferencias), no. Para un ser humano, si empieza a buscar el día 10 de enero, y encuentra una que se resuelve el 15 de septiembre, pero se encuentra cerrada desde el día 10 de diciembre del año anterior, se trata de una convocatoria a la que ya no se puede acceder, un ordenador puede creer que cumple la condición, etc.

Además, el grupo formado por ontología + sistema de inferencias + sistema de información debería ser capaz de superar lo que podemos denominar el *test google*, que dice lo siguiente: si una ontología no supera el rendimiento de una búsqueda en un motor convencional (p. e. *Google*) es que la ontología es deficiente (si la ontología no supera a un simple indizador de cadenas de caracteres, ¿para qué podríamos querer desarrollar ontologías con su enorme carga de trabajo?).

Por tanto, ¿pueden ser útiles las ontologías en sistemas de información? Por supuesto que sí, pero una ontología "sola" es, como dirían los americanos, un pato cojo.

4.3. Indeterminación

Por último, el problema más grave es que no existe un modelo consensuado y testado de forma amplia sobre cómo podría funcionar el triángulo Sistema de información – Sistema de inferencias – Ontología, sobre el cual nos centraremos ahora. El diagrama siguiente (figura 6) es una representación de lo que podríamos llamar el problema esencial de las ontologías ¿cómo se relacionan estos tres elementos?:

Aunque no faltan precisamente propuestas conceptuales y modelos de laboratorio, la industria aún no ha generado un modelo viable como el que ha establecido desde hace años con relación a los tesauros y los sistemas de información.

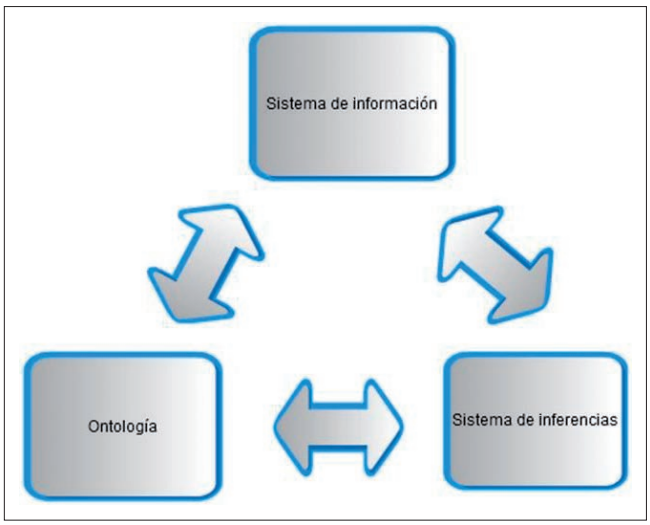

Figura 6. Triángulo esencial de las ontologías en relación con los sistemas de información

Probablemente no se ha conseguido, pese a lo prometedor de la tecnología, porque nadie sabe sobre seguro cómo podría funcionar ese triángulo. ¿Cómo actúa la ontología en relación a las preguntas del usuario? ¿Cómo actúa en relación con el fondo documental? ¿Dónde se sitúa el sistema de inferencias?

El problema de un modelo como el anterior es que no aclara nada en realidad a los ingenieros que deberían implementarlo, por no hablar de los costes de desarrollo. Peor aún, no aclara nada sobre cómo debería ser la interfaz de usuario de modo que fuera usable, que lo pudiera entender, interaccionar si el sistema no le ofrece lo que quería, etc. No olvidemos que los usuarios de sistemas de información odian la complejidad. Todo el proceso debería ser transparente, pero a la vez, inteligente, usando retroacción del usuario, etc.

4.4. Análisis de casos

En un loable intento de demostrar las capacidades reales, actuales y futuras de la web semántica (**Feigenbaum**, 2007), el *W3 Consortium* mantiene una interesante página (figura 7) donde recoge casos de aplicación práctica de la misma. Su dirección es:
http://www.w3.org/2001/sw/sweo/public/UseCases

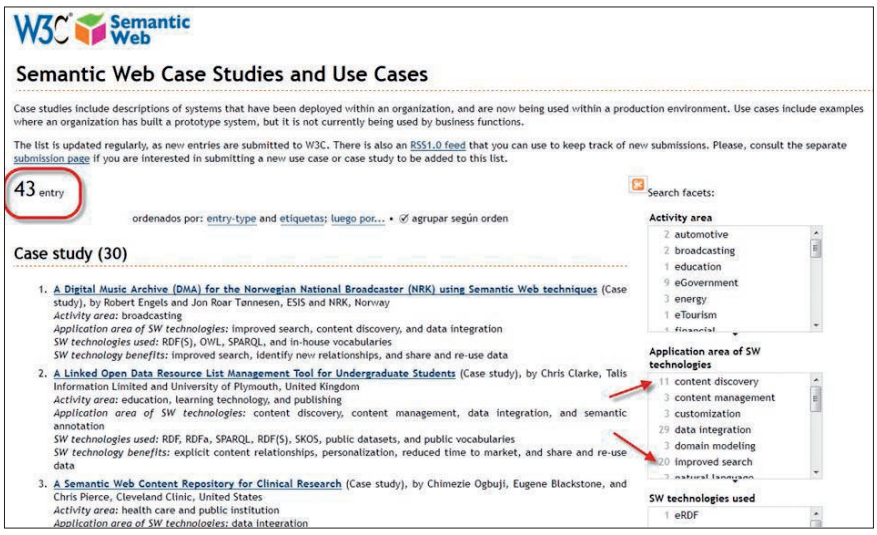

Figura 7. Página web con los casos de estudio del W3C (febrero 2011)

En ella podemos encontrar información sobre 30 *casos de estudio* y 13 *casos de uso*. La diferencia entre ellos radica en que mientras los casos de estudio describen implementaciones reales de las tecnologías de la web semántica, los casos de uso plantean sólo prototipos.

En cuanto a los dominios o sectores que llevan a cabo estas iniciativas (y atendiendo únicamente a los casos de estudio por ser de aplicación real), los desarrollos realizados tienen su origen principalmente en la industria TIC (Tecnologías de la Información y la Comunicación), aunque seguidos de cerca por la administración pública, el sector sanitario y los profesionales de biblioteconomía y documentación.

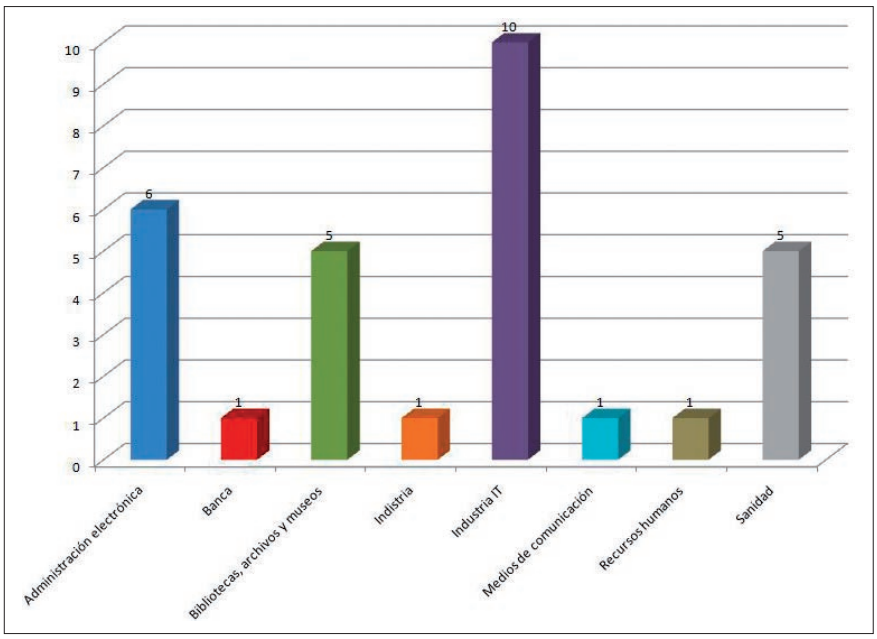

Figura 8. Histograma con el número de casos de estudio por sector del *W3C*

La figura 8 presenta un histograma que contabiliza el número de casos de estudio llevados a cabo por sector, y muestra el papel protagonista que han jugado bibliotecas, archivos y museos en la implementación real de las tecnologías de la web semántica (**Codina**, 2009).

En lo que a los ámbitos de aplicación de los casos de estudio se refiere, éstos son muy diversos si bien destaca de forma notoria el uso de estas tecnologías para la integración de datos (29% de los casos) y para la mejora de los sistemas de búsqueda utilizados en las instituciones desarrolladoras (21% de los casos).

Esta misma diversidad la encontramos en las tecnologías utilizadas, aunque con un claro predominio de aquellas que permiten la definición de lenguajes artificiales, concretamente: *RDF Schema* (utilizado en 24 casos), *OWL* (17 casos) y *SKOS* (3 casos).

Es interesante mencionar aquí que, el hecho de que *RDF Schema* (*W3C*, 2004b), que es un lenguaje para la descripción de vocabularios que permite principalmente la definición de *clases* o *tipos de cosas*, tenga un uso más extendido que los lenguajes *OWL* (para la definición de ontologías) (*W3C*, 2004a) o *SKOS* (para la formalización de tesauros) (*W3C*, 2009), así como que su uso aparezca vinculado en ocasiones a propuestas que mencionan la creación de ontologías (y en las que no se utiliza el lenguaje *OWL*), es un claro indicador de la imprecisión asociada al concepto de ontología.

Para terminar con el análisis de los casos de estudio, mencionar que la mayoría de ellos son interesantes, en especial algunos relativos a la *BBC*. Pero un examen detallado de los mismos arroja la siguiente casuística:

a) Un cierto número ya no están en activo. O sea, es imposible utilizarlos. Son servicios desconectados o desactivados o abandonados de algún modo.
b) Un cierto número, a pesar de estar anunciados desde hace años, aún no han entrado en funcionamiento, ni tienen fecha para ello.
c) Un cierto número se limitan (aunque es valioso en sí mismo, por supuesto) a utilizar sistemas de codificación de la web semántica en colecciones de datos, pero sin que suponga mejorar el sistema de información utilizando el triángulo esencial de la figura 6.

No cabe duda de que todos y cada uno de los casos supuso alguna mejora en los sistemas de información, pero apenas en el sentido fuerte del término, es decir, con un cambio de

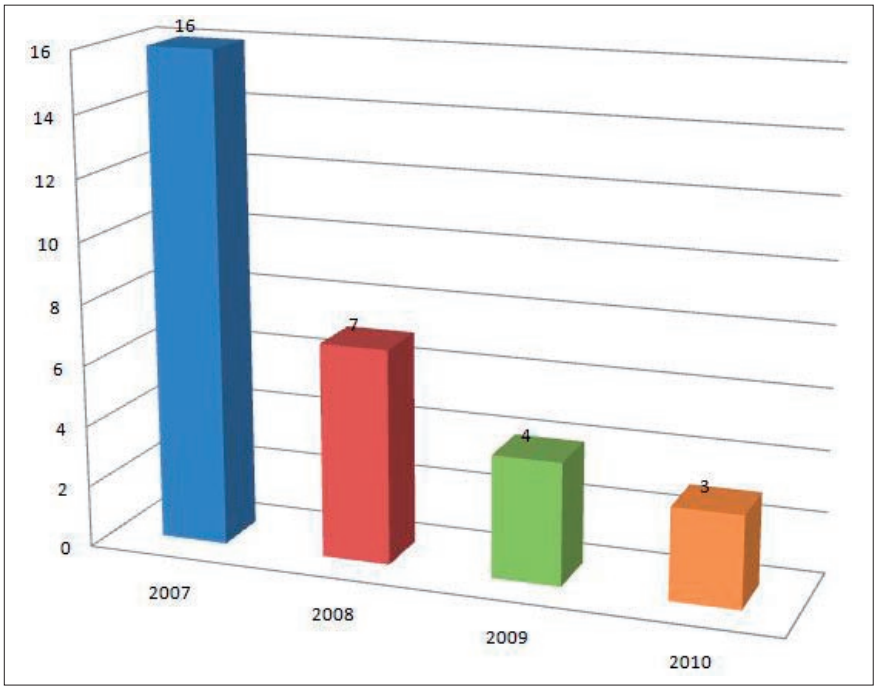

Figura 9. Evolución del número de casos de estudio del W3C (hasta finales de 2010)

paradigma como el que supuso en su momento la introducción de los tesauros en las bases de datos (comparado con las clasificaciones) o como un nuevo estándar para la industria de la información.

Además, aunque alguno de los casos suponga un éxito decidido, es muy significativo el pequeño número de casos de estudio (30) para un proyecto que data de 1998, y más aún cuando el análisis de estos casos de estudio (figura 9) revela una tendencia descendente en la aplicación de las tecnologías semánticas del *W3C*.

5. Conclusiones

Los tesauros son una tecnología (y también un *know-how*) bien asentada en los sistemas de información, con modelos bien establecidos y con una buena implantación en la industria y, por tanto, en la profesión.

Las ontologías son una tecnología prometedora y de un enorme potencial, ya probada en otros campos (por ejemplo, en la lingüística y en ámbitos específicos de la inteligencia artificial), pero en su aplicación a los sistemas de información aún es una solución inmadura.

No obstante los profesionales de la biblioteconomía y documentación debemos estar atentos a sus posibilidades. Sus fundamentos son sólidos y están atrayendo una gran cantidad de atención y energías. Es probable que en pocos años comiencen a arrojar resultados mucho más tangibles. Por el momento, han tenido el muy benéfico efecto de proporcionar sistemas de codificación y normas bien establecidas para la representación del conocimiento del que ya se ha beneficiado nuestra área como *SKOS* para representar tesauros mediante lenguajes propios de la web semántica (*resource description framework* o *RDF*).

Por supuesto, un tesauro representado en *SKOS* sigue siendo un tesauro, pero el hecho de disponer de un lenguaje de alto nivel formal y exigencia lógica como *RDF* puede ayudar no solamente a desarrollar mejores lenguajes documentales, sino que facilita la interconexión de entre los lenguajes desarrollados y especificados mediante *SKOS*. Además, adoptar las tecnologías de la web semántica, como las ontologías, pondrán las bases para la interconexión de datos y la prestación de mejores servicios de información.

Para finalizar, nos gustaría señalar que otra de las virtudes de las ontologías como campo de estudio es que está ayudando a revitalizar la semántica documental, tanto en la vertiente de los lenguajes documentales como en los esquemas de metadatos, dotando a ambos de una nueva e insospechada potencialidad.

Al mismo tiempo se están abriendo nuevas posibilidades para los profesionales y estudiosos que no duden en entrar en este nuevo territorio, todavía sin explorar totalmente, y por lo tanto con buenas perspectivas para hacer aportaciones. Para decirlo de otra forma, gracias a las ontologías y al proyecto de la web semántica, estamos de nuevo en una situación en la que es imposible enfocar el telescopio sin descubrir alguna cosa.

6. Referencias

6.1. Bibliografía

ANSI/NISO Z39.19-2005. *Guidelines for the construction, format, and management of monolingual controlled vocabularies*. Bethesda: NISO Press, 2005.
http://bit.ly/gUHNL)
http://www.niso.org/kst/reports/standards…

Arano, Silvia. "Los tesauros y las ontologías en la biblioteconomía y la documentación". *Hipertext.net*, 3, 2005.
http://www.upf.edu/hipertextnet/numero-3/tesauros.html

Baader, Franz; **Calvanese, Diego**; **McGuinness, Deborah L.**; **Nardi, Daniele**; **Patel-Schneider, Peter F.** *The description logic handbook: theory, implementation, applications*. Cambridge, UK: Cambridge University Press, 2003.

Bechhofer, Sean; **Goble, Carole**. "Thesaurus construction through knowledge representation". *Data & knowledge engineering*, 2001, v. 37, n. 2, pp. 25-45.

Berners-Lee, Tim; **Hendler, James**; **Lassila, Ora**. "The semantic web". *Scientific American*, 2001, v. 284, n. 5, pp. 34-43.

Centelles, Miquel. "Taxonomías para la categorización y la organización de la información en sitios web". *Hipertext.net*, 2005, 3.
http://www.upf.edu/hipertextnet/numero-3/taxonomias.html

Codina, Lluís; **Marcos, Mari-Carmen**; **Pedraza-Jiménez, Rafael**. *Web semántica y sistemas de información documental*. Gijón: Trea, 2009.

Feigenbaum, Lee; **Herman, Ivan**; **Hongsermeier, Tonya**; **Neumann, Eric**; **Stephens, Susie**. "The semantic web in action". *Scientific American*, 2007, v. 297, n. 6, pp. 90-97.

Fensel, Dieter; **Hendler, James A.**; **Lieberman, Henry**; **Wahlster, Wolfgang** (ed.) *Spinning the semantic web: Bringing the World Wide Web to its full potential*. Cambridge: The MIT Press, 2005.

García-Jiménez, Antonio. "Instrumentos de representación del conocimiento: tesauros versus ontologías". *Anales de documentación*, 2004, n. 7, pp. 79-95.
http://revistas.um.es/analesdoc/article/view/1691

Gil-Leiva, Isidoro. *Manual de indización: teoría y práctica*. Gijón: Trea, 2008.

Gilchrist, Alan. "Thesauri, taxonomies and ontologies, an etymological note". *Journal of documentation*, 2002, v. 59, n. 1, pp. 7-18.

Gruber, Thomas R. "A translation approach to portable ontologies". *Knowledge acquisition*, 1993, v. 5, n. 2, pp. 199-220.

Guzmán-Luna, Jaime A.; **Torres-Pardo, Durley**; **López-García, Alba-Nubia**. "Desarrollo de una ontología en el contexto de la web semántica a partir de un tesauro documental tradicional". *Revista interamericana de bibliotecología*, 2006, v. 29, n. 2, pp. 79-94.

"Lógicas de descripción". *Wikipedia*.
http://es.wikipedia.org/wiki/Lógica_de_descripción

López-Huertas, María-José. *Thesaurus structure design: a conceptual approach for improved interaction*. *Journal of documentation*, 1997, v. 53, n. 2, pp. 139-177.

Morales-del-Castillo, José M.; **Pedraza-Jimenez, Rafael**; **Ruiz-Rodríguez, Antonio A.**; **Peis-Redondo, Eduardo**; **Herrera-Viedma, Enrique**. "A semantic model of selective dissemination of information for digital libraries". *Information technology and libraries*, 2009, v. 28, n. 1, pp. 21-30.

Noy, Natalya F.; **McGuinness, Deborah L.** *Ontology development 101: A guide to creating your first ontology*. Stanford Knowledge Systems Laboratory, Technical report KSL-01-05, 2001.

Pollock, Jeffrey T. *Semantic web for dummies*. Hoboken, NJ: Wiley, 2009.

Pedraza-Jiménez, Rafael; **Codina, Lluís**; **Rovira, Cristòfol**. "Web semántica y ontologías en el procesamiento de la información documental". *El profesional de la información*, 2007, v. 16, n. 6, pp. 569-578.

Pérez-Agüera, José-Ramón. "Automatización de tesauros y su utilización en la web semántica. *BiD*, 2004, n. 13. *http://www.ub.es/bid/13perez2.htm*

Van-Slype, Georges. *Los lenguajes de indización: concepción, construcción y utilización en los sistemas documentales*. Madrid, Salamanca: Fundación Germán Sánchez Ruipérez, 1991.

Soergel, Dagobert. "The rise of ontologies or the reinvention of classification". *Journal of the American Society for Information Science*, 1999, v. 50, n. 12, pp. 1119-1120.

World Wide Web Consortium (W3C). *OWL web ontology language: overview* (*W3C recommendation*, 10 Feb 2004), 2004a. *http://www.w3.org/TR/owl-features*

World Wide Web Consortium (W3C). *RDF (resource description framework) vocabulary description language 1.0: RDF schema* (*W3C recommendation*, 10 Feb 2004), 2004b. *http://www.w3.org/TR/rdf-schema*

World Wide Web Consortium (W3C). *SKOS (simple knowledge organization system) primer: W3C Working group note*, 18 Aug 2009. *http://www.w3.org/TR/skos-primer*

6.2. Sitios de interés

Willpower. Thesaurus principles and practice. *http://www.willpowerinfo.co.uk/thesprin.htm*

W3C. Semantic web activity. *http://www.w3.org/2001/sw*

W3C - Semantic web use cases. *http://www.w3.org/2001/sw/sweo/public/UseCases*

Wikipedia. "Tesauros versus ontologías". *http://es.wikipedia.org/wiki/Tesauros_versus_ontologias*

7. Nota

Este trabajo forma parte del proyecto *Evolución de los ciberm*edios españoles en el marco de la convergencia. Análisis del mensaje, CSO2009-13713-C05-04, del *Ministerio de Ciencia e Innovación* (*Micinn*).