Noviembre-diciembre 2007, vol. 16, nº 6
ISSN 1386-6710 / e-ISSN 1699-2407
www.elprofesionaldelainformacion.com
Número dedicado a Ontologías

El profesional de la información

Revista internacional

Artículos

Ontologías

Observatorio

Análisis

El profesional de la información

Revista fundada en 1992 por Tomàs Baiget y Francisca García-Sicilia

REDACCIÓN:
El Profesional de la información
Apartado 32.280
08080 Barcelona
epi@elprofesionaldelainformacion.com

PUBLICIDAD:
Tel.: +34–609 352 954
publici@elprofesionaldelainformacion.com

SUSCRIPCIONES:
El profesional de la información
Apartado 32.280
08080 Barcelona, España
suscripciones@elprofesionaldelainformacion.com
http://www.elprofesionaldelainformacion.com/suscripciones.html

Teléfono de atención al suscriptor
+34 609 352 954

SERVICIOS ONLINE:
Catorze.Com
Apartado 41
08272 Sant Fruitós de Bages
Tel. +34-650 839 200
javier@catorze.com

DISEÑO:
Ignacio Pastor de la Huerta
ignacio@designio.es

MAQUETACIÓN:
Jorge Liras
Romargraf, S.A.

PRODUCCIÓN e IMPRESIÓN:
Romargraf, S.A.
Joventut, 55-57
08904 L'Hospitalet de Ll.
Tel. +34-933 345 466
romargraf@romargraf.es

DISTRIBUCIÓN ONLINE:
MetaPress, Alabama, EUA
http://elprofesionaldelainformacion.metapress.com

Depósito legal: B-12303-97

Los trabajos de la sección "Artículos" son aprobados según el sistema tradicional "peer review": al menos dos expertos en el tema, del consejo asesor de la revista y/o externos, deben dar el visto bueno antes de su publicación.

Para conseguir que los trabajos no pierdan su actualidad, la dirección y los evaluadores de esta revista ponen especial esfuerzo en revisar los artículos con gran rapidez, consiguiendo un tiempo medio de aceptación o rechazo de los trabajos de sólo unas pocas semanas.

Sumario

Tema central: Ontologías

Debido a la gran cantidad de originales recibidos sobre ontologías, algunos se publicarán en el siguiente número.

Los contenidos de **El profesional de la información** están referenciados en los siguientes servicios bibliográficos y bases de datos:

Bedoc
http://www.inforarea.es/bedoc.htm

Biblioteca de Andalucía
http://www.juntadeandalucia.es/cultura/b/cgi-bin/abweb/L1/T1/S09

Bire
http://gestiona.madrid.org/bire/servlet/Servidor?opcion=ConsultarGeneral&tipoBiblioteca=R&itBibliobuses=

Bulletin Board for Libraries (Bubl)
http://bubl.ac.uk/archive/journals/epdli/

Compludoc
http://www.ucm.es/BUCM/complu

ConnectSciences (Pascal)
http://connectsciences.inist.fr

Consorci de Biblioteques Universitàries de Catalunya (Cbuc)
http://sumaris.cbuc.es/13866710.htm

Datathéke
http://milano.usal.es/dtt.htm

Dialnet
http://dialnet.unirioja.es

DoIS (Documents in information science)
http://wotan.liu.edu/dois/data/julqtichq.html

E-LIS (E-prints in library and information science)
http://eprints.rclis.org/perl/search/advanced?=&publication=profesional+informacion

Ebscohost Electronic Journals Service
http://ejournals.ebsco.com/direct.asp?JournalID=105302

Gobierno de la Región de Murcia
http://www.carm.es/ctra/cendoc/bdatos/revistas/revista.asp?idrevista=290

Google Scholar
http://scholar.google.com

GVA (Generalitat Valenciana)
http://www.pre.gva.es/argos/va/contenido_general/recursos/bolsum/

Índice español de ciencias sociales y humanidades (ISOC)
http://bddoc.csic.es:8080/isoc.do

Information science and technology abstracts (Ista)
http://www.epnet.com/thisTopic.php?topicID=91&marketID=1

Information Services in Physics, Electronics and Computing (Inspec)
http://www.iee.org.uk/publish/inspec

IN-Recs
http://ec3.ugr.es/in-recs

Library and information science abstracts (Lisa)
http://www.csa.com/factsheets/lisa-set-c.php

Library, information science & technology abstracts (Lista)
http://www.libraryresearch.com

MetaPress
http://www.metapress.com/link.asp?id=105302

Oclc Firstsearch
http://www2.oclc.org/oclc/fseco/topic_area.asp?topic=Z

Registros Bibliográficos para Bibliotecas Públicas Españolas (Rebeca)
http://www.mcu.es/REBECA/que.html

Social science citation index (ISI, Social SCI)
http://go.isiproducts.com/

SwetsWise
http://www.swetswise.com/link/access_db?issn=1386-6710

Universidad de Castilla-La Mancha
http://biblioteca2.uclm.es/biblioteca/sumarios/pi.pdf

Universidad de Chile
http://www.al-dia.cl/sistema/tablas/listar.asp?r=3199

Universidad de Oviedo
http://librivision.uniovi.es/web/sumarios_web/Profesional-de-la-Informacion/

Ontologías y organización del conocimiento: retos y oportunidades para el profesional de la información

Por **Francisco-Javier García-Marco**

Resumen: *Se analiza la convergencia que se está produciendo en el campo de las ontologías entre ingeniería del conocimiento y organización del conocimiento en el marco del proyecto de la web semántica. Se estudia el desarrollo de la investigación sobre ontologías en las ciencias de la documentación y en el conjunto de las disciplinas que se interesan por los problemas ontológicos. Se contextualiza el actual frente de investigación en el campo de las ontologías en el marco del desarrollo de Internet y especialmente de la web semántica. Finalmente, se analizan las implicaciones de futuro para el profesional de la información: integración en un campo transdisciplinar más amplio y con un gran porvenir, clarificar su posición en él, y asegurar una formación adecuada en los nuevos estándares y tecnologías.*

Palabras clave*: Ontologías, Organización del conocimiento, Ingeniería del conocimiento, Internet, Web semántica.*

Francisco Javier García Marco *es doctor en filosofía y letras desde 1994 y profesor titular de universidad del área de biblioteconomía y documentación de la Universidad de Zaragoza desde 1996. Ha participado en la organización de numerosos eventos científicos, entre los que destacan los Encuentros Internacionales sobre Sistemas de Información y Documentación (Ibersid), de los que es director. Dirige las revistas Scire: representación y organización del conocimiento e Ibersid.*

Title: Onthologies and knowledge organization: challenges and opportunities for information professionals

Abstract*: The emerging convergence in the field of ontologies between knowledge organization and knowledge engineering is examined in the context of the semantic web project. We describe the development of research on ontologies in library and information sciences and in other disciplines interested in ontological problems. The emergence of a research agenda on ontologies is discussed in the context of the development of the internet and, specifically, in relation to the semantic web project. Finally, some implications of this convergence for the information professional are discussed: the integration of knowledge organization into a broader transdisciplinary research arena with a very promising future, the need to clarify the role and potential contributions of the information professional, and the urgency of adequate education and training in the new standards and technologies.*

Keywords*: Ontologies, Knowledge organization, Knowledge engineering, Internet, Semantic web.*

NOTA: *Este artículo ha sido evaluado y aprobado por pares.*

García-Marco, Francisco-Javier. "Ontologías y organización del conocimiento: retos y oportunidades para el profesional de la información". En: *El profesional de la información,* 2007, noviembre-diciembre, v. 16, n. 6, pp. 541-550.

DOI: 10.3145/epi.2007.nov.01

1. Las ontologías: un campo de convergencia transdisciplinar

EN LOS ÚLTIMOS AÑOS EXISTE UNA FUERTE tendencia en el campo de las ciencias de la documentación por asimilar los sistemas de organización del conocimiento tradicionales –clasificaciones bibliográficas y archivísticas, tesauros de distintos tipos, lenguajes de encabezamientos de materias y de autoridades, principalmente– y los que han surgido en el nuevo entorno de internet –taxonomías corporativas y de usuario, términos libres asignados (*tags*), etc.– a un concepto nuevo procedente de la inteligencia artificial: las ontologías (v. g. Vickery, 1997; Baeza; Ribeiro, 1999; Currás, 2005; Hersh, 2005).

Se trata de una situación semejante a lo que ocurre en el campo de la descripción bibliográfica y documental en general, que cada vez se percibe más claramente

Artículo recibido el 13-07-07
Aceptación definitiva: 20-09-07

integrado con el movimiento general de descripción de objetos informáticos mediante metadatos. Por otra parte existe la tendencia paralela de los informáticos que trabajan en aplicaciones de lenguajes documentales en internet así como de los gestores que llevan y promueven los proyectos a denominarlos también ontologías. Como resultado se está produciendo una rápida convergencia de ambos lenguajes.

¿Por qué sucede esto? Sin duda existen varias razones. Una de ellas es el hecho obvio de que una de las aplicaciones más poderosas de las ontologías –inicialmente desarrolladas para soportar la inferencia[1] lógica en sistemas expertos y de adquisición de conocimiento– es la recuperación de información. Otra condición importante es que los lenguajes de descripción de ontologías formales implementados sobre software –por ejemplo, *Protégé*– permiten fácilmente representar los objetos, las relaciones y restricciones existentes en los lenguajes documentales, y facilitar su interoperabilidad en el marco de los sistemas de información interconectados.

La familia de términos "ontolog-" no es ajena a la tradición de los lenguajes documentales; antes al contrario, ha estado muy presente a través del problema de las categorías universales de la clasificación; hasta el punto en que autores como **Dahlberg** (1978) han hablado tempranamente de ellas como de "estructuras ónticas". Además, la indudable modernidad y prestigio que rodea a los términos procedentes de la informática avanzada –y en general de los campos de investigación punteros– favorece su adopción. Por último, el acceso a la información en la Web ha topado finalmente con el problema conceptual, y existen cada vez más voces –entre ellas la de su fundador **Tim Berners-Lee** (1998, 2001)– que reclaman el abordaje del problema del significado –el problema *semántico*– para que la información de la Red se pueda relacionar y utilizar de forma más precisa y potente. Siendo como es internet la tecnología de nuestros tiempos, la asunción del problema semántico dentro de ella sitúa la organización del conocimiento en un punto privilegiado, aunque subsidiario. Todo ello ha favorecido un proceso de rápida asimilación conceptual y terminológica, y el término ontologías se utiliza cada vez con más profusión.

Sin embargo, se olvida frecuentemente que dichas ontologías "informáticas" no son lenguajes documentales al modo tradicional, sino implementaciones conjuntas de sistemas de términos sobre diversas lógicas –generalmente, la de predicados de primer orden–, siendo lo auténticamente distintivo lo segundo. Esas herramientas sirven para la representación del conocimiento de cara a soportar las inferencias de sistemas expertos. Sus aplicaciones transcienden la recuperación de información, que constituirían dentro de ellas –nada más, pero nada menos– que una subdisciplina específica.

"El problema semántico en internet sitúa la organización del conocimiento en algo privilegiado"

Por otra parte, y aunque se interesen en ellos, la especialidad de las ontologías formales "informáticas" no aborda sistemáticamente ni agota la totalidad de los problemas ontológicos ni en documentación ni en otras disciplinas adyacentes, especialmente el de las categorías universales que son del interés de la ontología filosófica, la semántica y la organización del conocimiento; o las leyes de la estructuración del significado que estudia la semántica. Las ontologías aportan una formalización de las meta-relaciones ontológicas que permite las operaciones lógicas entre ellas, pero la especificación de los entes, las restricciones y las relaciones específicas queda como un problema dentro de cada dominio y aun a nivel universal en buena parte. Esto es, las ontologías "informáticas" han implementado los aspectos ontológicos que contiene la lógica de primer orden, pero los aspectos categoriales quedan en gran parte fuera de la formalización (**Poli**, 2002).

Se vislumbra en ocasiones el peligro de confundir los diferentes instrumentos de organización conceptual y terminológica, haciendo sinónimos conceptos próximos pero que son claramente diferentes –sistemas de organización del conocimiento, tesauros, taxonomías y ontologías-. Esta posición es comprensible desde la perspectiva gerencial, general y lejana de los responsables máximos de las servicios de información, y de las personas interesadas en el tema. Sin embargo es inaceptable entre los expertos y, especialmente, dentro del campo de investigación, donde la asimilación conceptual y terminológica sólo contribuye a crear confusión. Este problema afortunadamente es uno de los que queda meridiamente desentrañado en la nueva norma británica *BS 8723* (*British Standards Institute*, 2006-). En este aspecto cabe ser optimistas: está claro que, superada la confusión inicial, los diferentes conceptos ocuparán el lugar que les corresponde.

Por tanto, parece que estamos ante la aparición de uno de esos nuevos campos de investigación eminentemente transdisciplinares como las ciencias cognitivas, la neurociencia o las ciencias ambientales en los que ninguna disciplina termina por organizar a las demás, ni tampoco es posible obviar los avances que se producen en los diferentes campos. Ciencias ontológicas es, posiblemente, un término presuntuoso, porque, aunque variado e interdisciplinar, el número de investigadores

activos y trabajos publicados en el área, como se verá seguidamente, es muy reducido frente a las grandes macrodisciplinas que se acaban de mencionar. En cualquier caso, la expansión de los términos metadatos y ontologías en las ciencias de la documentación constituye la punta del iceberg de un proceso de reconfiguración disciplinar, como resultado de la integración de diversas ciencias en el marco de desarrollo tecnológico info-comunicacional (figura 2).

"Una ontología es un sistema de términos que sirve para describir y representar un área de conocimiento y que expresa las relaciones entre ellos por medio de un lenguaje formal que puede ser entendido por un ordenador"

2. Advenimiento de la tecnología ontológica y el desarrollo de internet

2.1. Caracterizando el campo de las ontologías técnicas

Ahora bien, ¿qué son esas ontologías que han irrumpido en la investigación moderna en las ciencias de la computación?

Uno de los más famosos investigadores y pionero en el tema, **Tom Gruber** (1993) las define simplemente como una "especificación de una conceptualización"; mientras que, más recientemente, la *International DOI Foundation* (2005) define una ontología estructurada como "an explicit formal specification of how to represent the entities that are assumed to exist in some area of interest and the relationships that hold among them", es decir, como una especificación formal explícita –una declaración– sobre la manera de representar las entidades existentes en un área de interés y las relaciones que mantienen entre ellas. Ésta y otras definiciones, sin embargo, son muy amplias para el no experto y pueden servir para describir realidades de un nivel de complejidad muy diferente: listas de palabras (vocabularios simples), jerarquías de términos (taxonomías) o complejas formalizaciones léxicas como *WordNet 2.0*, que es capaz de representar relaciones de hiperonimia[2], hiponimia[3], instanciación[4], así como otras relaciones léxicas, incluyendo la representación morfológica. Por otra parte, no diferencian adecuadamente lo que aportan de nuevo las ontologías frente a las clasificaciones, los tesauros y otros lenguajes documentales.

Lo primero que conviene establecer es que las ontologías son un campo de investigación de la inteligencia artificial y más específicamente de la rama relacionada con la representación del conocimiento, la ingeniería del conocimiento, que se ocupa de la construcción de sistemas expertos. Se trata de un área de investigación que en su época emergente –y todavía hoy– ha tenido un carácter marcadamente interdisciplinar con aportaciones de la filosofía, la lingüística y las ciencias cognitivas en general (**Nicles**; **Pease**; **Schalley**; **Zaefferer**, 2007).

El objetivo de la ingeniería del conocimiento es constituir grandes bases de conocimientos sobre un tema en forma de declaraciones, reglas de inferencia y mecanismos de razonamiento (motor de inferencia) para resolver automáticamente problemas del dominio en cuestión. Si se quiere, y dicho más sencillo, para responder automáticamente preguntas sobre el dominio de representación, ya sea a un agente humano, a otro automático, con el fin de ayudarle en su proceso de toma de decisiones y, eventualmente, de ejecución de una tarea. Las ontologías son un procedimiento basado en la lógica de primer orden desarrollado para codificar adecuadamente el sistema de términos utilizados en dichas declaraciones, de forma que se exprese adecuadamente la relación entre los términos; esto es, son herramientas para construir sistemas conceptuales o, por utilizar una terminología común, vocabularios estructurados, en los que se explicitan todas las relaciones entre los términos que se utilizan y otras restricciones de significado.

McGuinness (2002) realiza una descripción de sus características: deben poseer un vocabulario controlado limitado –aunque extensible–, con una interpretación estricta de sus clases y relaciones entre términos –sin ambigüedades– y una relación jerárquica estricta entre sus clases; como propiedades típicas aunque no obligatorias, deben permitir la especificación de propiedades para cada clase, la inclusión de individuos –ejemplares– en la ontología y la especificación de restricciones de valor a nivel de cada clase; y, como propiedades deseables –ni obligatorias ni típicas–, se recomienda que permitan la especificación de clases disjuntas, la especificación de relaciones arbitrarias lógicas entre los términos y la distinción de relaciones, como inversas y parte-todo.

El objetivo de las ontologías es constituir un almacén de información semántica donde sea posible consultar el significado de un término a través de los mecanismos proposicionales propuestos ya por Aristóteles en su teoría de la definición (1965, ed.). Los significados de los términos son resueltos por los expertos en ontologías a partir de las relaciones entre los mismos: "an ontology is a document or file that formally defines the relations among terms" (**Berners-Lee**; **Hendler**; **Lassila**, 2001).

¿Por qué hacerlo? Ciertamente, dichos significados se pueden dar por obvios en un sistema experto cerra-

do; pero es absolutamente necesario declararlos explícitamente para intercambiar información entre diferentes sistemas expertos que se comunican entre sí. Como dice **Gruber** (1993), "lo importante es para qué sirve una ontología. Mis colegas y yo hemos estado diseñando ontologías con el objetivo de facilitar el intercambio y la reutilización de conocimientos".

Imaginemos los sistemas de información de dos empresas de trabajo temporal de dos países distintos –por ejemplo, Reino Unido y España– intercambiando datos de ofertas y demandas de trabajo para equilibrar sus respectivos mercados. El sistema de una empresa inglesa puede solicitar a un servidor español empleados con un determinado perfil si existe un vocabulario que exprese la relación interlingüística entre los respectivos términos. Ese vocabulario puede residir en cualquiera de los dos sistemas o, aún mejor, ser independiente de los mismos, de manera que otros que requieran una información semejante puedan aprovechar esa información.

En definitiva, una ontología es un sistema de términos que sirve para describir y representar un área de conocimiento, y que expresa las relaciones entre ellos por medio de un lenguaje formal (lógico) que puede ser entendido por un ordenador. Comparte el vocabulario y su estructuración con los lenguajes de descriptores, los tesauros y las taxonomías que se utilizan para la organización de la información, que utilizan conjuntos de términos relacionados para describir un dominio o área de conocimiento. Sin embargo, implica una descripción formal exigente de esas relaciones que pueda ser interpretada por un ordenador. Por ello, gran parte del esfuerzo en el campo de las ontologías ha sido dedicado a elaborar lenguajes capaces de ello.

Existen varios lenguajes disponibles. Algunos han sido desarrollados en el ámbito de la inteligencia artificial como *CycL*, el lenguaje del proyecto *Cyc* (**Stephen**; **Lenat**, 2002), *Ontolingua* (2005) o el *Knowledge Interchange Format* (*KIF*) y su sucesor el *Simplified common logic*, o la *Knowledge representation system specification* (*KRSS*) (**Patel-Schneider**; **Swartout**), orientados a la representación de enunciados en la lógica de primer orden. Sin embargo, el más importante ahora es el *Web ontology language* (*OWL*) desarrollado sobre rdf y rdfs, y deudor a su vez de *OIL*, *DAML* y *DAML+OIL*.

2.2. Las ontologías en la www

Las ontologías han recibido un impulso decisivo de la estrategia de web semántica del *World Wide Web Consortium* (*W3C*) (1994-2004). Como es bien conocido, el objetivo es llevar la web a su máximo potencial, es decir, al máximo nivel de automatización en los procesos de transferencia de la información y el conocimiento. Hasta ahora la www ha facilitado enormemente el proceso de compartir información entre personas gracias a sus eficaces estándares de comunicación –http– y de normalización de los documentos –html–. Sin embargo no permite la recuperación y procesamiento de la información a nivel de dato y combinaciones de ellos (información), sino tan sólo de documento, lo que es imprescindible para los procesos de automatización que soportan los diferentes tipos de lenguajes de programación. Los líderes del programa lo expresan así: "To date, the web has developed most rapidly as a medium of documents for people rather than for data and information that can be processed automatically. The semantic web aims to make up for this" (**Berners-Lee**; **Hendler**; **Lassila**, 2001).

Así por ejemplo, en la actualidad es posible consultar las páginas web con ofertas de viaje a un destino dado, pero es imposible saber sin leerlas cuáles son más convenientes. Actualmente esto sólo podría hacerse después de un cuidadoso análisis y de su introducción en una base de datos, pero eso requiere una operación centralizada o una cuidadosa concertación entre una red de bases de datos.

¿Cuál es la alternativa que ofrece la web semántica? Propone etiquetar esos documentos de ofertas de manera que ciertas informaciones como los lugares de salida, destino, medio, clase y precio pudieran ser fácilmente reconocibles, y que luego un programa pudiera recopilar esos datos, procesarlos y presentar resultados al usuario. Más aún, dichas etiquetas podrían incluso ser distintas siempre que en algún lugar de la red estuviera disponible una equivalencia de las mismas y de sus relaciones. Por fin, sus contenidos podrían ser procesados si se almacenasen en algún lugar común las relaciones entre ellos, que es de lo que se ocupan las ontologías (que también tienen un gran potencial para ocuparse del problema anterior). Así, por ejemplo, si se produce una oferta genérica de vuelos a Brasil con una tarifa común a todos los destinos internos llegando a Río de Janeiro, sería posible que el sistema valorara esa oferta automáticamente en el caso de que quisiéramos ir a Brasilia, si un fichero especificara que esta ciudad existe y es una subclase de Brasil.

En definitiva, como dice la web oficial (*World Wide Web Consortium*, 1994-2004): la web semántica trata de dos cosas. Se ocupa de que existan formatos comunes de intercambio de datos, mientras que en la web original sólo teníamos intercambio de documentos. Además tiene que ver con la elaboración de un lenguaje para codificar cómo los datos se relacionan con los objetos del mundo real. Eso permitiría a una persona o a una máquina comenzar a trabajar en una base de datos, y luego moverse a través de otras que no están conectadas por cables, sino por tratar del mismo asunto.

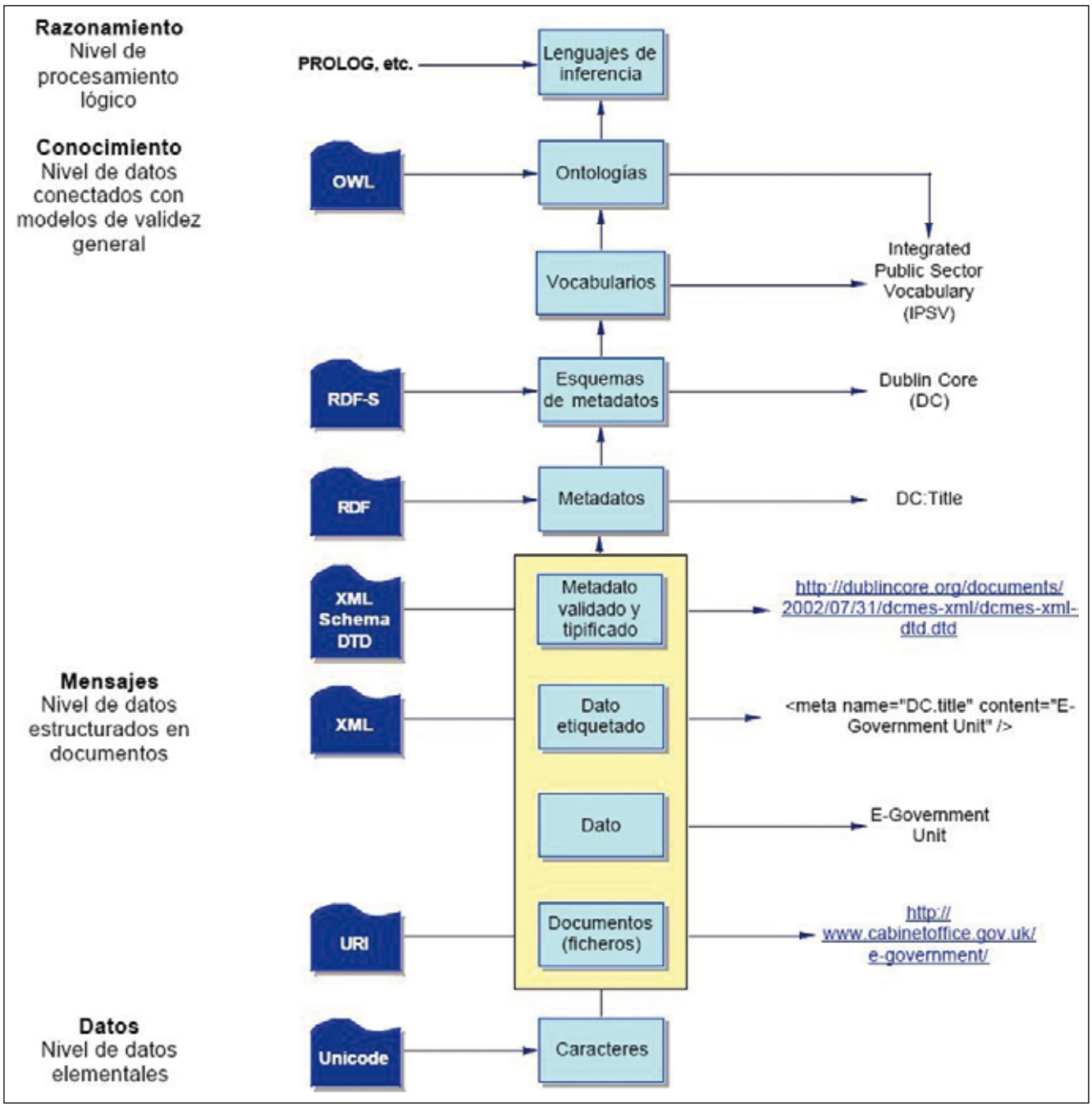


Figura 1. La arquitectura de la web semántica y los procesos cognitivos (adaptación de García-Marco, 2006)

La figura 1 resume el conjunto de estándares que van a convertir la www en un entorno de intercambio no sólo de documentos, sino también de datos y de relaciones entre los mismos (**García-Marco**, 2006)

Como se puede apreciar, la última capa de estándares está formada por el *OWL*, que constituye el estándar propuesto por el *World Wide Web Consortium* (2004) para la descripción formal de ontologías. Por fin, existe un estándar *ISO* que se puede utilizar para la presentación de ontologías de cara a la recuperación de la información y a la presentación de las relaciones para la navegación: se trata de los mapas temáticos o *topic maps* (*International Organization for Standardization*, 1998).

3. Mirando al futuro

Es fácil caracterizar el mundo actual desde el punto de vista informacional: internet se ha convertido en el nuevo entorno de distribución, almacenamiento, publicación y acceso a la información. El abaratamiento que ha conllevado en los procesos de información en todas sus fases ha sido enorme, y como consecuencia, tanto la información disponible como la accedida crecen a un ritmo vertiginoso. El proceso de digitalización que se está realizando con el objetivo de trasladar todos los contenidos a la Red es imparable y, como resultado, los que antes se distribuían por los diferentes canales de comunicación existentes –prensa y publicaciones periódicas, libro, cine, televisión, radio, dvd, audio cd,

etc.– se están trasladando al nuevo medio masivamente y cada vez con más rapidez. Por todo ello parece claro que los esfuerzos de la organización del conocimiento como disciplina –y de la documentación en general (v. g. **López-Yepes**, 1998, p. 17)– deben concentrarse en el nuevo medio, sin que eso deba suponer, lógicamente, descuido de los demás.

Por otra parte, la descomunal sobreabundancia de información que conlleva internet está produciendo un viraje decisivo en la manera de entender la gestión de la información: el acento que tradicionalmente se ponía en la conservación –del lado de los custodios– y en la velocidad y potencia de la recuperación –desde el punto de vista de los tecnólogos– ha cedido a un nuevo énfasis en la selección y el filtrado (**García-Marco** 2002). Esta nueva corriente no debería suponer en ningún caso descuido ni desprecio de las funciones de conservación y recuperación masiva, pero resulta absolutamente necesario en el nuevo contexto.

"La web semántica es a la www lo que los tesauros fueron a la recuperación en texto libre en los albores de la documentación automatizada"

Todo ello redunda en una renovada vigencia de las herramientas de control conceptual y terminológico –que, como sabemos, en el campo de la biblioteconomía y documentación se han venido denominando lenguajes documentales-. La primera impresión es que esa situación privilegiada se refuerza al converger con el movimiento de las ontologías. En cierta manera, y en el campo específico de la recuperación de la información, la web semántica es a la www lo que los tesauros fueron a la recuperación en texto libre en los albores de la documentación automatizada. También entonces la fascinación inicial por la búsqueda mediante palabras extraídas cedió enseguida ante la necesidad del control terminológico, lo que llevó a su vez a la invención y a la generalización de los tesauros (**Moreiro**, 2007), que finalmente convergieron con las clasificaciones bajo el paraguas del movimiento analítico. De forma semejante, el proyecto de la web semántica pretende añadir a la www lo que le falta en el ámbito de la recuperación y el procesamiento de la información: granularidad[5], precisión, exhaustividad e inferenciabilidad en el manejo a la información. Con estos nuevos objetivos, tanto los ingenieros del conocimiento como los profesionales de la información y la documentación ven multiplicada su relevancia y sus posibilidades de acción.

En este contexto, el despegue de la investigación sobre ontologías y el movimiento de convergencia de la investigación sobre organización del conocimiento con la que se realiza sobre ontologías aparecen llenos de oportunidades –y también de alguna amenaza–. La convergencia es tanto más interesante cuanto que, por un lado, se apoya en un sustrato común –la tradición ontológica y lógica, y el paradigma del procesamiento de

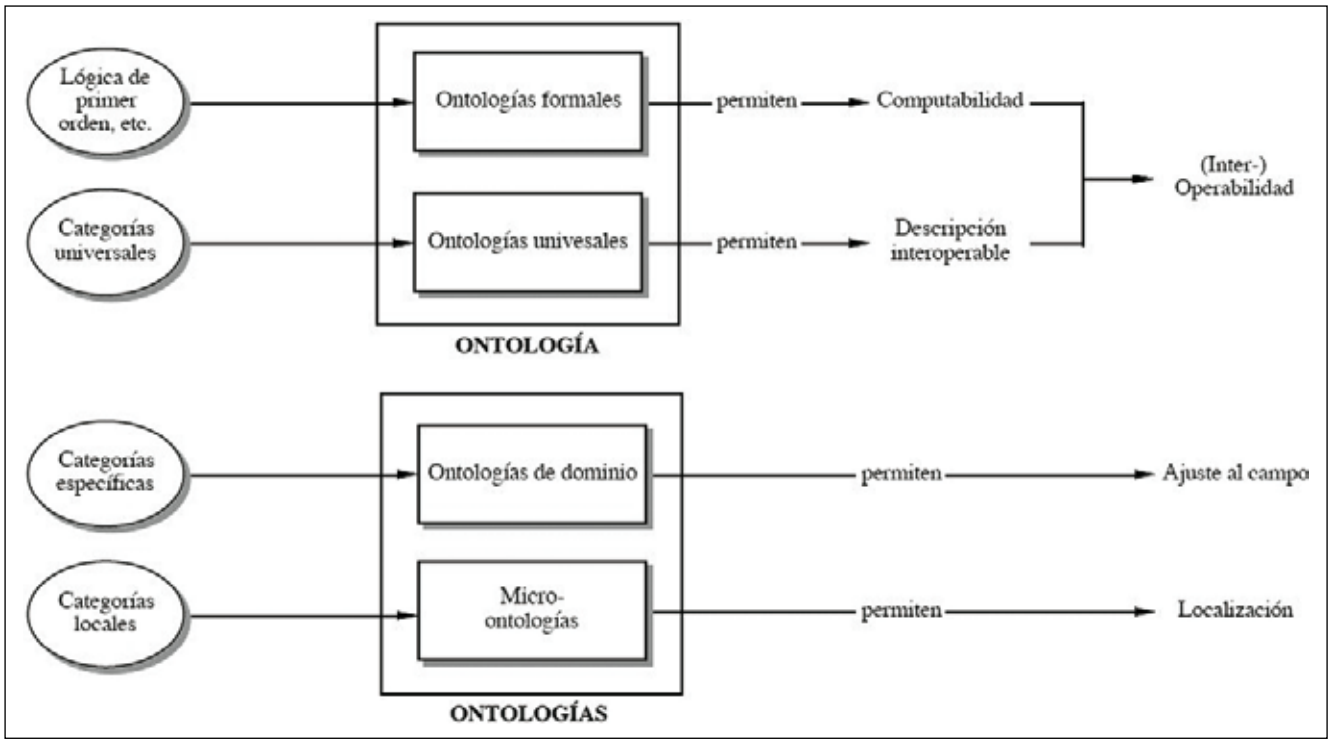


Figura 2. La ecología de las disciplinas ontológicas

la información—; y, por el otro, sucede de la mano de un conjunto de ciencias con las que ambas vienen históricamente manteniendo una relación estrecha (figura 2). De esta manera, la organización del conocimiento se inserta en un campo más amplio de investigación transdisciplinar –¿ciencias ontológicas?– que, por otra parte, cuenta entre sus fuerzas con un interesantísimo, potente y vital frente de investigación.

En el plano práctico, la convergencia entre sistemas de organización del conocimiento y ontologías promete dar abundantes frutos. La implementación de lenguajes documentales en forma de ontologías facilita la realización de operaciones lógicas sobre ellos: expansiones y restricciones de búsquedas, desambiguaciones, etc. Ya era posible en sistemas automatizados anteriores, pero de forma "cerrada". La asunción de estándares abiertos en el entorno de la web semántica permite hacer todo lo anterior a una escala anteriormente impensable. Al compartir las diferentes ontologías es posible una "gramática" común, por ejemplo, detectar un término en diversas fuentes de información, comprobar sus términos genéricos y equivalencias, desambiguarlo, unir hipónimos de diferentes ontologías y presentarlos en un orden dado al usuario o resolver directamente los recursos a los que apuntan.

Los cambios raramente vienen tan sólo con oportunidades y en este caso, como en tantos otros, también hay que salir al paso de amenazas. Los peligros más importantes que plantea el proceso de convergencia a la organización del conocimiento y a la documentología son, por un lado, el ser arrastrado a la confusión inicial que rodea a todo este magma investigador, y, por el otro, el de ser absorbido y perder su entidad. La solución a todo ello pasa por clarificar campos de práctica, entender el lenguaje de las otras disciplinas, explicar el propio, interconectar –pero también deslindar– los programas de investigación, y determinar con precisión las aportaciones y tareas más importantes de unos y de otros.

En esta dirección, la figura 3 intenta bosquejar un deslinde de la I+D+I en el campo de las ontologías que lleva implícito una propuesta de división del trabajo. Se detectan cuatro ámbitos distintos, agrupados a su vez en dos bloques. El superior está constituido por "la ontología" propiamente dicha, en singular, que se ocuparía de estudiar y desarrollar los aspectos generales comunes a todas las implementaciones ontológicas con el fin de asegurar su interoperabilidad global, necesaria en internet. Dentro de ella, se identifican dos zonas diferentes: a) la superior, que se refiere a la formalización de las ontologías en lenguajes lógico-matemáticos de cara a asegurar su computabilidad y, en definitiva, la automatización de las tareas ontológicas de propósito general; y b) la inferior, que se refiere a la teoría de la descripción de los entes en el nivel máximo de generalidad, y que tiene como tarea fundamentalmente el estudio de las categorías universales.

"Los cambios raramente vienen tan sólo con oportunidades y, en este caso, como en tantos otros, también hay que salir al paso de amenazas"

El bloque inferior se dedica a "las ontologías" en plural; esto es, a las concreciones de los modelos generales en los diferentes dominios específicos (disciplinas, subdisciplinas, etc.). Este bloque se concibe también dividido en dos capas: a) una referente a la disciplina o área disciplinar específica (las ontologías de dominio); y b) otra que se ocupa de los aspectos idiosincrásicos de una determinada comunidad o entidad (micro-ontologías). En esta arquitectura existen muchos nichos para las diferentes disciplinas de la figura 2, notablemente, la investigación sobre las categorías universales de organización del conocimiento, el lenguaje y los documentos; las propiedades formales de las ontologías; o las implementaciones tecnológicas. Sin embargo, los resultados en cualquiera de estas áreas serán mejores en la medida en que se reconozca la unidad del área y se realice un auténtico trabajo transdisciplinar.

Además del posicionamiento dentro de este esquema, el investigador y profesional de la información tiene ante sí la tarea de abordar tres frentes diferentes: el de la teoría, el de la práctica y el de la formación. En el primero urge integrar la reflexión ontológica realizada en la organización del conocimiento con la que se ha realizado en las tecnologías informáticas y la inteligencia artificial, más cercana a la filosofía analítica dentro de la inteligencia artificial y sus aplicaciones; y seguir con atención y cuidado los desarrollos que se producen en ese campo y en las otras disciplinas relacionadas para no perder el paso y quedar descolgados. En una segunda fase habrá que comunicar en el nuevo lenguaje aquellos aspectos propios en los que hay mucho que aportar, ahorrando a los colegas esfuerzos de "reinventar la rueda".

En el frente de la práctica, es necesario aprender a expresar los diferentes lenguajes documentales en los nuevos formalismos y tecnologías, y desde esa experiencia y esa práctica colaborar en la tarea común: el desarrollo de vocabularios estructurados para la recuperación de información en el contexto de internet y sus aplicaciones corporativas (**García-Marco** et al, 2007).

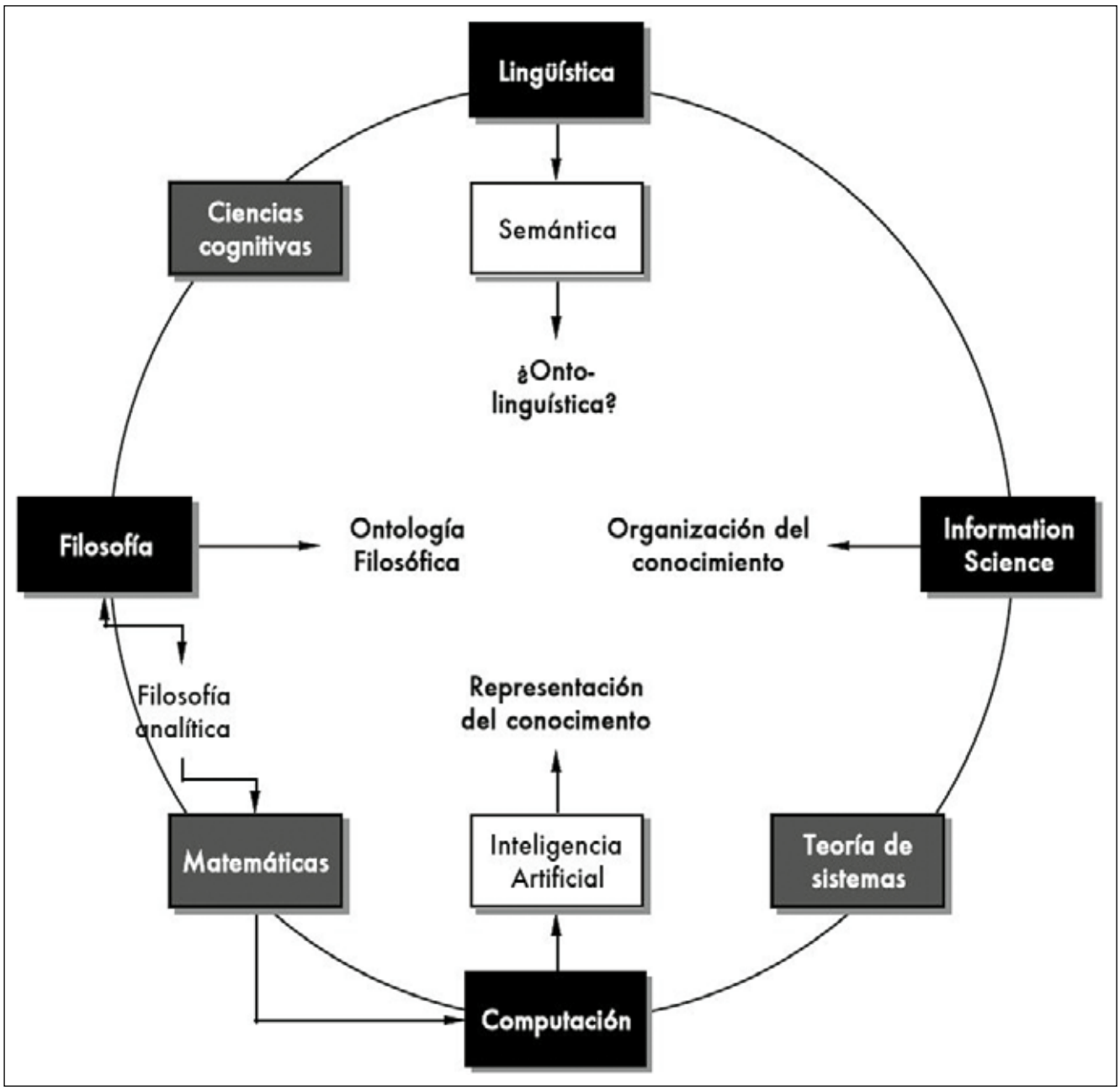


Figura 3. Ontología y ontologías

"Ante las ontologías, el investigador y profesional de la información debe abordar tres frentes diferentes: teoría, práctica y formación"

Por último, es necesario insistir en que para hacer posibles los objetivos anteriores, se debe realizar un importante esfuerzo en el campo de la formación inicial y continua, pues el éxito requiere que los profesionales conozcan y sean capaces de utilizar los estándares de representación y navegación ontológica, y especialmente las nuevas herramientas de gestión de ontologías y de utilización de las mismas en entornos de recuperación de información; o al menos, que sean capaces de trabajar eficaz y eficientemente con las personas que se encarguen de estos menesteres. Esto no será posible sin reforzar, –siquiera mínimamente y con carácter instrumental– la formación lógico-matemática de los futuros profesionales de la información, una preocupación sentida y abordada ya por **Sayers** (1926), en la que ha insistido en nuestro país **Izquierdo** (1990) y que ha sido recientemente revisada por **Bufrem**, **Breda** y **Sorbías** (2007).

Teoría, práctica, formación y apertura interdisciplinar son, en definitiva, los indicadores que van a marcar el éxito en la nueva oportunidad que las ontologías y la web semántica abren al profesional e investigador de la información y la documentación.

Notas de la Redacción

1. Inferencia es el acto o proceso de derivar una conclusión basada solamente en lo que uno ya conoce.
Se usa en varios campos:
La inferencia humana (o sea, cómo los humanos sacan conclusiones) se estudia tradicionalmente en psicología cognitiva. La Lógica estudia las leyes de la inferencia válida.
Los estadísticos han desarrollado reglas formales de inferencia de datos cuantitativos. Los investigadores en inteligencia artificial realizan sistemas automáticos de inferencia.
Fuente: Wikipedia inglesa

2. En semántica lingüística se denomina hiperónimo a aquel término general que puede ser utilizado para referirse a la realidad nombrada por un término más particular.
Semánticamente, un hiperónimo no posee ningún rasgo semántico, o *sema*, que no comparta su hipónimo, mientras que éste sí posee rasgos semánticos que lo diferencian de aquél.
Por ejemplo, coche posee sólo los semas [+vehículo], [+con motor] y [+pequeño tamaño], que comparte con descapotable, mientras que descapotable posee además el rasgo [+sin capota], que lo diferencia de coche.
Al redactar un texto conviene utilizar hiperónimos para evitar la repetición de palabras ya empleadas anteriormente, como se hace en el siguiente ejemplo: "De repente, un *descapotable* rojo paró frente al banco. Del *automóvil* salieron dos individuos encapuchados, mientras otro esperaba en el *vehículo*".
Fuente: Wikipedia española

3. En lingüística, un hipónimo es una palabra o frase cuya gama semántica está incluida dentro de otra palabra. Por ejemplo, escarlata, bermellón, carmín y carmesí son hipónimos de rojo (su hiperónimo).
Según **Victoria Fromkin** y **Robert Rodman** ("Introducción al lenguaje", ISBN 0-03-018682-X), hipónimos son un conjunto de palabras relacionadas cuyo significado son los casos específicos de una palabra más general (así, por ejemplo, rojo, blanco, azul, etc., son hipónimos de color). La hiponimia es pues la relación entre un término general tal como el *polígono* y los casos específicos, como *triángulo*.
Fuente: Wikipedia inglesa

4. Instanciación. Veamos unos conceptos:
Clase. Define los aspectos abstractos de una cosa (objeto), incluyendo sus características (atributos, campos de aplicación o propiedades) y sus conductas (las que puede hacer, sus métodos o su comportamiento). Por ejemplo, la clase "perro" consistiría en rasgos compartidos por todos perros, como la especie (características), y la habilidad de ladrar (la conducta).
Objeto. Es una *instancia* (un caso) particular de una clase. La clase "perro" define todos los perros posibles listando sus características y sus conductas; el objeto "Lassie" es un perro concreto, con versiones particulares de las características. Un "perro" tiene piel; "Lassie" tiene piel marrón y blanca. El objeto "Lassie" es una *instancia* (o sea, un caso particular) de la clase "perro". El conjunto de valores de los atributos de un objeto particular se llama su estado. El objeto consiste del estado y del comportamiento que se ha definido en la clase del objeto.
Fuente: Wikipedia inglesa

5. La granularidad es una medida del tamaño de los componentes de un sistema. La granularidad es el tamaño relativo, la escala, el nivel de detalle o profundidad de la penetración que caracteriza un objeto o una actividad. Es "el grado en que se subdivide una entidad mayor". Por ejemplo, un metro desmenuzado en milímetros tiene la granularidad más fina que un metro desmenuzado en centímetros".
Fuente: Wikipedia inglesa

Referencias bibliográficas

Aristóteles. *Categorías*. Traducción del griego y prólogo de **Francisco de P. Samaranch**. 2a ed. Buenos Aires: Aguilar, 1965.

Baeza Yates R.; **Ribeiro Neto, B.** *Modern information retrieval.* Harlow, England: Addison Wesley, 1999.

Berners-Lee, T.; **Hendler, J.**; **Lassila, O.** "The semantic web". En: *Scientific Americ*an, 2001, v. 284, n. 5, pp. 34-43.

Berners-Lee, Tim. *Semantic web road map: an attempt to give a high-level plan of the architecture of the semantic www*. W3C, September 14-10-94. Consultado en: 12-07-07.
http://www.w3.org/DesignIssues/Semantic.html

British Standards Institute. *BS 8723, Structured vocabularies for information retrieval*. London: British Standards Institute, 2006-.

Bufrem, Líela Santiago; **Breda, Sônia Maria**; **Sorbías, Tidra Viana**. "The presence of logic in the domain of knowledge organization: interdisciplinary aspects of college curricula". En: **Rodríguez Bravo, Blanca**; **Alvite Díez, María Luisa** (eds.). *La interdisciplinariedad y la transdisciplinariedad en la organización del conocimiento científico: Actas del VIII Congreso ISKO-España, 2007*. León: Universidad de León, Secretariado de Publicaciones, pp. 179-185.

Currás, Emilia. *Ontologías, taxonomía y tesauros: manual de construcción y uso*. 3ª ed., act. y amp. Gijón: Trea, 2005.

Dahlberg, I. *Ontical structures and universal classification*. Bangalore, Sarada Ranganathan Endowment for Library Science, 1978.

García-Marco, Francisco-Javier. "La literatura científica sobre lenguajes poscoordinados en España: de la divulgación del concepto a la internet". En: *Documentación de las ciencias de la información*, 2002, v. 25, pp. 291-319.

García-Marco, Francisco-Javier. "Ontologías y documentación electrónica en las actividades públicas". En: **Galindo, Fernando** (ed.). *Gobierno, derecho y tecnología: las actividades de los poderes públicos*. Madrid: Thomson-Civitas, 2006, pp. 173-225.

García-Marco, Francisco-Javier (coord.); **Agustín Lacruz, Carmen**; **Caro Castro, Carmen**; **Martínez Usero, José Ángel**; **San Segundo, Rosa**. "Proyectos internacionales de reforma y ampliación de las normas sobre tesauros para su adaptación a los nuevos contextos de integración e interoperabilidad en el entorno digital". En: **Rodríguez Bravo, Blanca**; **Alvite Díez, María Luisa** (eds.). *La interdisciplinariedad y la transdisciplinariedad en la organización del conocimiento científico: Actas del VIII Congreso ISKO-España, 2007*. León: Universidad de León, Secretariado de Publicaciones, pp. 27-40.

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

Hersh, William Richard. "Ontologies for information retrieval". En: **Jorde, L. B.**; **Little, P.**; **Dunn M.**; **Subramaniam, S.** (eds.). *Encyclopedia of genetics, genomics, proteomics and bioinformatics*. London: John Wiley & Sons, Part 4, 2005.

International DOI Foundation. *The DOI Handbook, Version 4.2.0, released February 2005*. Oxford: International DOI Foundation (IDF), 2005. Consultado en: 12-07-07.
http://www.doi.org/handbook_2000/glossary.html

International Organization for Standardization. *ISO/IEC 13250, Information Technology-SGML Applications-Topic Maps*. Geneva: ISO, 1998.

Izquierdo Arroyo, José María. *Esquemas de lingüística documental*. Barcelona: PPU, 1990.

López Yepes, José. "Hombre y documento: del homo sapiens al homo documentator". En: *Scire*, 1998, julio-diciembre, v. 4, n. 2, pp. 11-22.

McGuinness, D. L. "Ontologies come of age". En: **Fensel, D.**; **Hendler, J.**; **Lieberman, H.**; **Wahlster, W.** (eds.). *Spinning the semantic web: bringing the world wide web to its full potential*. Cambridge: MIT Press, 2002.

Moreiro, José Antonio. "Evolución paralela de los lenguajes documentales y la terminología". En: **Rodríguez Bravo, Blanca**; **Alvite Díez, María Luisa** (eds.). *La interdisciplinariedad y la transdisciplinariedad en la organización del conocimiento científico: Actas del VIII Congreso ISKO-España, 2007*. León: Universidad de León, Secretariado de Publicaciones, pp. 27-40.

Nicles, Matthias; **Pease, Adam**; **Schalley, Andrea C.**; **Zaefferer, Diezmar**. "Ontologies across disciplines". En: **Schalley, Andrea C.**; **Zaefferer, Dietmar**. *Ontolinguistics: how ontological status shapes the linguistic coding of concepts*. Berlin, etc.: Mouton de Gruyter, 2007, pp. 23-67.

Ontolingua. Standford: Stanford University, 2005. Consultado en: 12-07-07.
http://www.ksl.stanford.edu/software/ontolingua/

Patel-Schneider P. F.; **Swartout, B.** "Description-logic knowledge representation system specification". En: *KRSS group of the ARPA knowledge sharing effort*. Consultado en: 12-07-07.
http://www-db.research.bell-labs.com/user/pfps/papers/krss-spec.ps

Poli, R. "Glanzing at the problems of contemporary ontology". En: *Scire*, 2002, enero-junio, v. 8, n. 1, pp. 17-40.

Sayers, W. C. **Berwick**. *A manual of classification for librarians & bibliographers*. London: Grafton, 1926.

Schalley, Andrea C.; **Zaefferer, Diezmar**. *Ontolinguistics: how ontological status shapes the linguistic coding of concepts*. Berlin, etc.: Mouton de Gruyter, 2007.

Stephen, R.; **Lenat, D.** "Mapping ontologies into Cyc". En: *AAAI 2002 conference workshop on ontologies for the semantic web*, 2002.

Vickery, B. C. "Ontologies". En: *Journal of information science*, 1997, v. 23, n. 4, pp. 277-286.

World Wide Web Consortium (1994-2004). *Semantic web activity.* Cambridge, Keio, Paris: W3C, 1994-2004. Consultado en: 12-07-07. *http://www.w3.org/2001/sw/*

World Wide Web Consortium (2004). *OWL. Web Ontology Language Overview. W3C Recommendation 10 February 2004*. Cambridge, Keio, Paris: World Wide Web Consortium, 2004. Consultado en: 12-07-2007. *http://www.w3.org/TR/2004/REC-owl-features-20040210/*

***Francisco-Javier García-Marco**, Área de biblioteconomía y documentación, Universidad de Zaragoza.*
jgarcia@unizar.es

Lenguajes documentales y ontologías

Por **Rodrigo Sánchez-Jiménez** y **Blanca Gil-Urdiciain**

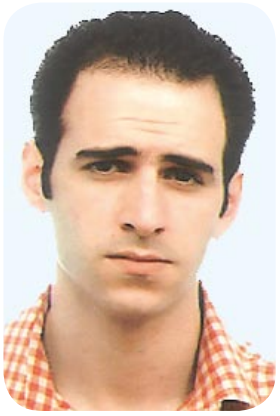
Resumen: *Este artículo analiza los principales puntos de convergencia y divergencia entre los lenguajes documentales y las ontologías en tanto que herramientas para la organización del conocimiento y para la recuperación de información en el ámbito de la web semántica. Se describen los aspectos fundamentales de una ontología, así como las principales características semánticas y estructurales de los lenguajes documentales para establecer una comparación entre ellos.*

Palabras clave: *Sistemas de clasificación, Tesauros, Ontologías, Web semántica, Relaciones semánticas, Análisis comparativo.*

Title: Controlled indexing languages and ontologies

Abstract*: The main points of convergence and divergence between controlled indexing languages and ontologies, tools for knowledge organization and information retrieval in the semantic web, are described. Fundamental aspects of ontologies are presented, as well as the basic semantic and structural characteristics of traditional controlled languages.*

Keywords*: Subject schemes, Thesauri, Ontologies, Semantic web, Semantic relationships, Comparative analysis.*



Rodrigo Sánchez Jiménez *es profesor en la Facultad de Ciencias de la Información de la Universidad Complutense de Madrid. Ha participado en proyectos de investigación relacionados con lenguajes documentales, gestión de documentación audiovisual y recuperación de información.*

Blanca Gil Urdiciain *es profesora en la Facultad de Ciencias de la Documentación de la Universidad Complutense de Madrid. Ha dirigido proyectos de investigación en el campo del análisis documental, indización y lenguajes documentales.*

Sánchez Jiménez, Rodrigo; Gil Urdiciain, Blanca. "Lenguajes documentales y ontologías". En: *El profesional de la información,* 2007, noviembre-diciembre, v. 16, n. 6, pp. 551-560.

DOI: 10.3145/epi.2007.nov.02

1. Introducción

Tim Berners-Lee (1999) define la web semántica como "una extensión de la web actual en la que se proporciona a la información un significado bien definido, lo que permite a la gente y las computadoras trabajar en cooperación"[1]. Por tanto, se trata de una nueva concepción de la www en la que el significado de las cosas y la capacidad de hacerlo inteligible a las máquinas juegan un papel esencial. Para hacerlo posible se cuenta con la capacidad de rdf para establecer descripciones de recursos mediante un modelo sencillo pero potente. Rdf ofrece la posibilidad de que distintas aplicaciones intercambien esta información en condiciones óptimas que permitan su reutilización sin que se pierda significado en el proceso, o sea, que la inicialmente dispuesta para una aplicación o un dominio en concreto pueda ser aplicada en un ámbito distinto al original[2].

El modelo de descripción de recursos de rdf se basa en grafos que se pueden expresar como sentencias (oraciones), en las que el sujeto es el recurso a describir, el predicado es una propiedad o característica propia de dicho recurso y el objeto es el valor concreto que tiene dicha característica. Una sentencia rdf cualquiera tiene por tanto una forma básica como la mostrada en la figura 1. Ambos nodos del grafo (sujeto y objeto) pueden formar parte de otras descripciones, de modo que varias descripciones se pueden unir en un solo grafo descriptivo más denso y con información más rica (figura 2).

El grafo anterior describe las relaciones entre una persona y el sitio web que ha creado, su contribución

Artículo recibido el 13-07-07
Aceptación definitiva: 21-09-07

Figura 1. Grafo representando una sentencia rdf básica

en la revista *Cuadernos de documentación multimedia*, en la que se publicó el artículo identificado por el *URI*:

http://multidoc.rediris.es/cdm/viewarticle.php?id=43

Esto establece una red de relaciones utilizables para diferentes propósitos de forma sencilla.

Sin embargo, el significado de cada uno de los atributos (los arcos o "flechas" del grafo), así como de los tipos de recursos (por ejemplo el tipo "persona"), se debe definir en otro lugar. Es decir, cada grafo rdf establece tan sólo descripciones de recursos de forma similar a una oración más o menos compleja. No obstante, el significado preciso de las palabras que forman dicha oración se debe definir fuera de él (o nos veríamos obligados a concretar cada elemento en juego cada vez que lo utilizamos).

La definición de las "palabras" que se pueden utilizar en una descripción rdf se hace en "vocabularios", mediante *schemata* rdf. La utilización de un *schema* nos permite definir el vocabulario propio de un dominio para las descripciones rdf, o sea, hace posible delimitar cualidades útiles para describir recursos, así como tipos de recursos de forma explícita para que puedan ser empleados en razonamientos automáticos.

Por otra parte, la utilización de esta codificación de las entidades y propiedades típicas de un dominio debería ser reutilizable fuera del mismo, con el objeto de crear una red de significados interrelacionada, legible y utilizable por máquinas. Es para este propósito para el que las ontologías tienen una gran importancia. No sólo nos permitirán definir vocabularios de forma más precisa, sino que admitirán la posibilidad de relacionar entre sí varios de ellos pertenecientes a diferentes dominios.

Recientemente se ha planteado como un tema de investigación de primera línea la relación entre las ontologías y los lenguajes documentales y se han llevado a cabo numerosos estudios comparativos (**Miles**, 2001; **Guzmán Luna**, et. al., 2006; **Soergel**, et. al., 2004). Además se ha discutido mucho sobre el papel que las ontologías podrían jugar como un nivel superior de desarrollo de los lenguajes documentales.

2. Ontologías

Existen varios paradigmas y muchas formas concretas de representar del conocimiento, pero las ontologías parecen ser la mejor forma en el ámbito de la web semántica. El *W3C* está haciendo un esfuerzo bastante importante en la difusión del lenguaje para la creación de ontologías *OWL* (*Ontology web language*), tal y como se expresa en el *Semantic web activity statement* (**Miller**, 2004) y en la necesidad de crear ontologías con las que expresar formalmente el significado de las cosas. Su utilización es clave desde el punto de vista de la reutilización del conocimiento en contextos diferentes al original, ya que por su estructura y capacidad de formalización permiten la puesta en relación de diferentes *schemata* rdf.

Una ontología, tal y como se entiende el término en filosofía, es un registro sistemático de las cosas que existen. Esta idea fue tomada del campo de la inteligencia artificial con algunas modificaciones, de forma que para un sistema lo que existe es lo que puede ser representado (**Noy**; **McGuinness**, 2000), lo que conlleva que una ontología será un registro de lo que puede ser representado.

"Existen varios paradigmas y formas concretas de representación del conocimiento, pero las ontologías parecen ser lo mejor para la web semántica"

Si buscamos una definición formal de ontología nos encontraremos con que existen varias posibilidades, por lo que optaremos por la clásica ofrecida por **Noy** y **McGuinness** (2000), según la cual "(...) es una descripción formal y explícita de los dominios del discurso"[3]. **Gruber** (1993) nos ofrece otra más abstracta: "especificación explícita de una conceptualización" siendo ésta "una visión abstracta y simplificada del mundo que queremos representar con algún propósito"[4]. Una conceptualización se compondría a su vez de objetos, conceptos y otras entidades que existen en un

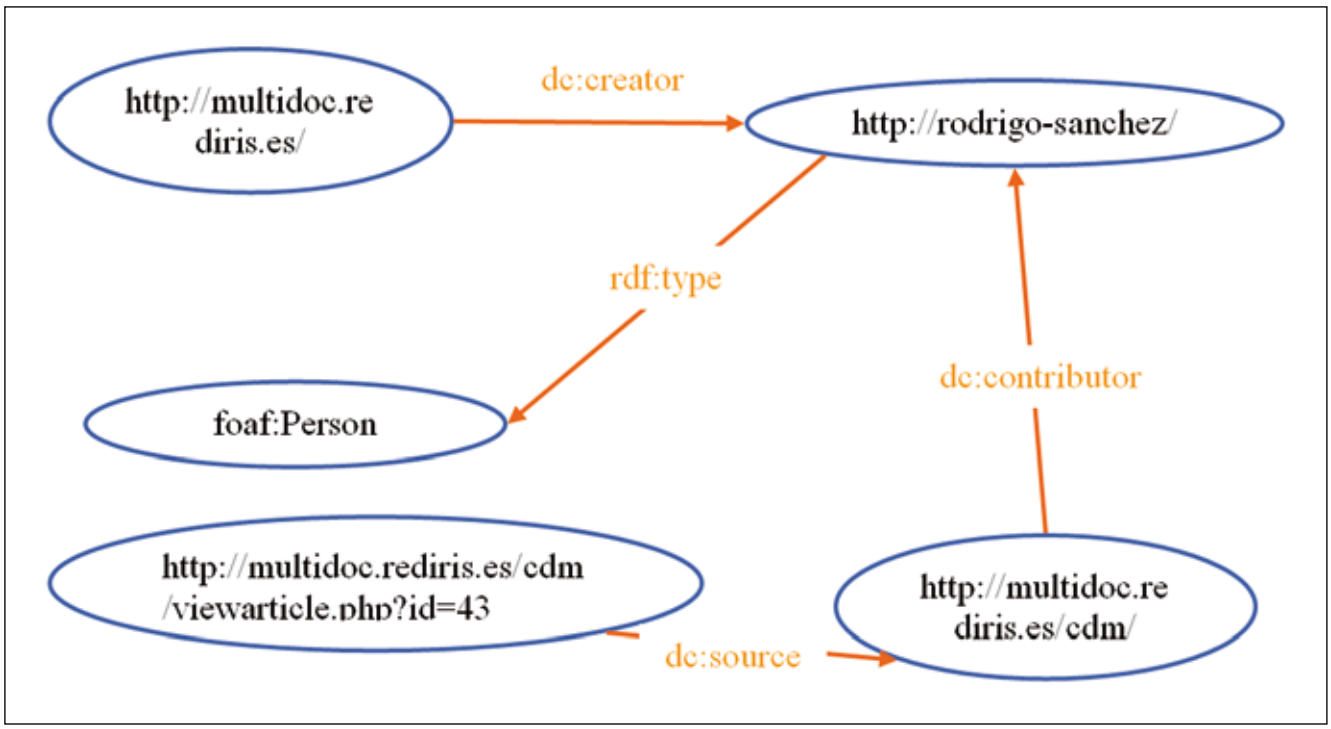


Figura 2. Grafo rdf representando múltiples relaciones sujeto/objeto

área de interés (o dominio) y las relaciones entre ambos. Existen muchas otras definiciones que se recogen en el trabajo de **García Jiménez** (2004).

Lo fundamental de la combinación de los elementos antes mencionados es la posibilidad de formular descripciones utilizables por ordenadores acerca de conceptos básicos en un dominio concreto, así como las relaciones entre ellos, lo que permite reutilizar el conocimiento entre varias áreas de conocimiento. Más concretamente **Noy** y **McGuinness** (2000) destacan cinco aspectos de utilidad:

– Compartir una comprensión común de la estructura de la información entre personas y agentes de software.

– Permitir la reutilización del conocimiento propio de un dominio.

– Hacer las asunciones propias de un dominio explícitas para terceras partes ajenas a él.

– Separar el conocimiento del dominio del conocimiento operacional.

– Analizar el conocimiento del dominio.

Las ontologías se han incorporado con fuerza al ámbito de la web (**Heflin**, 2004) y son utilizadas con cierta frecuencia en la esfera comercial[5]. En general se puede decir que se usan por personas, bases de datos o aplicaciones que necesitan compartir información acerca de un dominio concreto. Algunos ejemplos son los servicios avanzados de búsqueda conceptual, ayuda en la toma de decisiones, entendimiento del habla y el lenguaje natural, gestión del conocimiento, comercio electrónico, etc.

Es muy interesante, para el caso que nos ocupa, recoger la distinción entre ontologías ligeras y pesadas expresada por **Corcho**, **Fernández** y **Gómez** (2003). Las primeras incluyen conceptos, taxonomías de conceptos, relaciones entre conceptos y propiedades que los describen, mientras que las pesadas tienen en cuenta además axiomas y restricciones de las propiedades. Algunos autores, como **García Jiménez** (2004), vinculan las ontologías ligeras a los tesauros; en nuestra opinión no se puede hablar de una asimilación total debido a la inexistencia de una formalización de las clases en los lenguajes documentales, además de algunos otros aspectos, como se verá más adelante.

Esta distinción entre ontologías ligeras y pesadas parece recogerse en el ámbito de la web semántica mediante la especificación de varios sublenguajes de *OWL* (**McGuinness**; **Van Harmelen**, 2004). La web semántica requiere ontologías con un alto grado de estructuración (**Heflin**, 2004) que especifiquen la descripción de, al menos:

– Clases, tipos de cosas propias de un dominio de interés.

– Relaciones entre esas cosas.

– Sus propiedades o atributos.

Las clases en una ontología se definen a través del conjunto de atributos que comparten los miembros (o instancias) de dicha clase; es decir, existe una formalización explícita de cada una de ellas. Cada uno de los

atributos de una clase expresa una característica, cuyo valor puede ser restringido o matizado, acotando dominios (las clases sobre las que se puede aplicar una propiedad) y rangos (las clases a las que debe pertenecer el valor de una propiedad), cardinalidad, etc.

Podemos establecer vínculos entre clases, al margen de las relaciones de herencia que las vinculan con sus sub-clases, mediante la utilización de propiedades con rangos o dominios relativos a otras clases, de forma que un atributo deba tener como valor un miembro de una determinada clase.

El aspecto externo de una ontología es bastante similar al de un sistema de clasificación, como se puede apreciar en la figura 3. En realidad, los lenguajes documentales permiten establecer relaciones entre conceptos, que pueden ser en principio similares a las que ofrece una ontología, aunque cubriendo eso sí un espectro menos amplio. Sin embargo, un análisis más detallado de las características de los lenguajes documentales y las ontologías muestra que existen diferencias importantes entre ambos.

3. Diferencias entre ontologías y lenguajes documentales

Son apreciables las divergencias tanto a nivel semántico como estructural. En el primer caso, ya sean los sistemas de clasificación o las listas de encabezamientos de materias, se limitan a representar mediante términos la información contenida en los documentos. Los tesauros, en cambio, han tratado de representar mediante conceptos dicha información, con la ayuda de un sistema de relaciones que, aunque complejo, no alcanza la capacidad descriptiva que se desarrolla en las ontologías.

En este sentido, podemos decir que los tesauros intentan llegar al nivel conceptual a través de la utilización de relaciones que están fuertemente ancladas a nivel léxico. Es decir, que se interpreta que el nivel conceptual representado por los términos que componen un tesauro está directamente asociado a dichos términos. Las redes de asociaciones entre términos nos llevan a un término que actúa como representante único de un determinado concepto, pero su asociación con él está ahí desde el principio, anclada en la relación entre el léxico y el significado. En el caso de las ontologías el significado de las cosas se explicita a través de atributos, de las características que son propias de dichas cosas, y no de una representación léxica de las mismas.

Creemos que una de las mejores formas de observar las diferencias entre los lenguajes documentales y las ontologías en el nivel semántico es analizar con detenimiento el concepto de clase, el cual es primordial para entender el funcionamiento de las ontologías, y también un elemento crucial desde la perspectiva de la representación del conocimiento.

3.1. En torno al concepto de clase

Maniez (1992) lo define, en el caso de los lenguajes documentales como: "un conjunto de objetos (en el sentido amplio de la palabra) que tienen al menos un carácter en común". Por nuestra parte, según definíamos en un trabajo anterior (**Gil**, 2004), se puede entender como: "grupo de objetos o asuntos que comparten una o más características, identificadas normalmente por una notación específica". En el ámbito de las ontologías (**Bechhofer**, et. al., 2004) se define como: "un mecanismo de abstracción para agrupar recursos con similares características".

Sin embargo, al margen de las definiciones del concepto, existen algunas diferencias prácticas; una de ellas hace referencia al hecho de que en el ámbito de las ontologías una clase se define por un conjunto de propiedades, o atributos, que tendrán valores distintos en diferentes instancias de dicha clase y se declararán de forma explícita.

En el ámbito de los lenguajes documentales una clase efectivamente hace referencia a un conjunto de elementos con características similares, pero que no se explicitan en lugar alguno, salvo quizá en forma de notas explicativas, lo que en cualquier caso no es suficiente para su tratamiento automatizado. El problema de esta definición es que no se formaliza, es decir, no se asigna un conjunto de atributos que deben estar presentes en una entidad cualquiera para que pueda ser considerada miembro de una clase. Esto reduce mucho la capacidad de llevar a cabo razonamientos sobre la estructura del lenguaje documental, ya que las clases no se codifican de forma que resulten explícitas para una máquina, por lo que el conocimiento inherente en cada una de las clases queda implícito y no es utilizable.

En otras palabras, es fundamental que una ontología cumpla bien su papel de base de conocimiento para un ámbito en concreto. Para que sea así, debe ser capaz de responder una serie de preguntas a través del conocimiento codificado en ella, denominadas "preguntas de competencia" (**Noy**; **McGuinness**, 2000), que en el caso de una ontología que modelara una universidad, podrían ser las siguientes:

– ¿Es el decano de una facultad responsable de las infraestructura de redes?

– Las clases de la licenciatura de documentación ¿se imparten en la facultad de documentación o en la de ciencias de la información?

– ¿Dónde debo presentar mis impresos de matriculación?

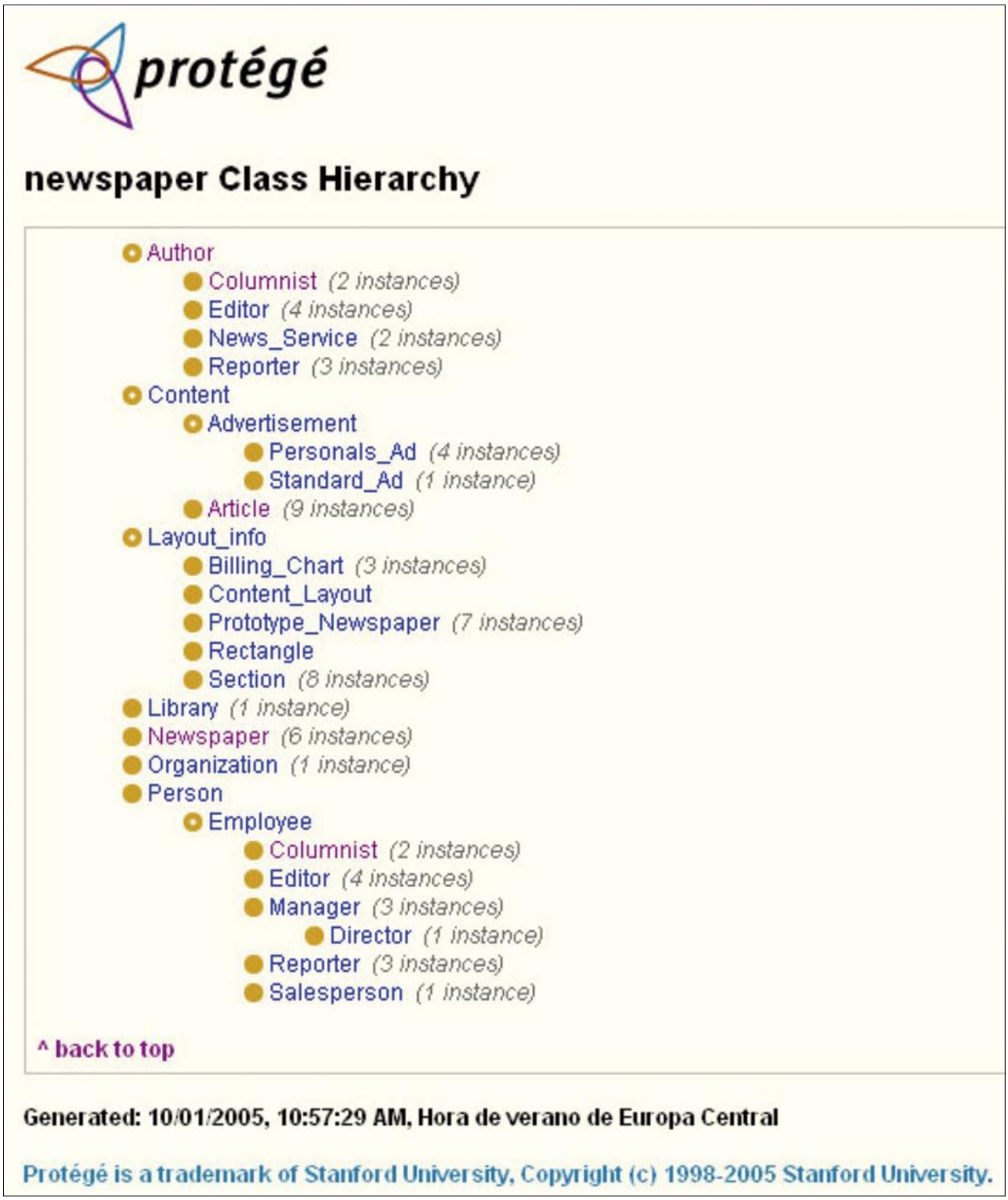


Figura 3. Presentación de la jerarquía de clases de una ontología. Fuente: Protégé (Universidad de Stanford)

Sin embargo, la mera existencia de una clase no proporciona la información necesaria para responder ese tipo de preguntas y por lo tanto no podría cumplir adecuadamente su papel. Para que esto sea posible es indispensable definir propiedades para cada una de las clases.

Las clases en un lenguaje documental, por ejemplo, en un sistema de clasificación, sí poseen atributos, expresados en forma de relaciones con otras clases y propiedades con valores determinados. El problema es que dichos atributos se aplican de forma indistinta a todas las clases del sistema de clasificación, o dicho de otra forma, sólo existe una clase formalmente diferenciada en un sistema de clasificación, la "clase concepto". **Miles** (2005) defiende la existencia de una "capa de indirección" entre los lenguajes documentales y las rea-

lidades que modelan. Hace referencia al hecho de que una entrada de un sistema de clasificación o de un tesauro representa un concepto útil para la descripción de documentos, pero no se constituye en el equivalente en la realidad de dicho concepto. Esto se encuentra directamente relacionado con el hecho de que los lenguajes documentales surgen para normalizar e internacionalizar conceptos que a su vez se utilizan para representar entidades del mundo real. Podemos utilizar los ejemplos del propio **Miles** para esclarecer este aspecto.

El grafo de la figura 4 nos muestra cómo el concepto "Enrique VIII", identificado por el URI: *http://www.example.org/concepts#henry8* (del tipo *skos: Concept*) y los atributos "fecha de creación" y "creador" se pueden aplicar únicamente al concepto generado para representar a la persona de Enrique VIII y no evidentemente a la propia persona de Enrique VIII. Podemos observar cómo se genera de forma natural una clase que describe un tipo de entidades con características comunes (los conceptos de un lenguaje documental). Sin embargo, no se genera una clase para cada una de las entidades del mundo real que vamos a representar, de manera que los atributos de Enrique VIII como instancia de la clase persona no se formalizan y no podemos llevar a cabo ninguna tarea que implique la utilización del concepto Enrique VIII en tanto que persona, sino que tendremos que manejar dicha entidad a través de un mecanismo indirecto, mediante los conceptos que el lenguaje documental utiliza para representarla.

En resumen, podemos decir que la única clase existente en un lenguaje documental, al menos desde el punto de vista de la metodología para crear ontologías, es la clase "concepto", que tendría a su vez como subclases materia, encabezamiento de materia o descriptor, cada una de las cuáles comparte una serie de características básicas con su superclase.

En el caso de un sistema de clasificación o de una lista de encabezamientos de materias, cada una de las entradas existentes puede ser entendida a su vez como una instancia de una única clase, la de materia, y los únicos atributos serían los de "notación", nota de alcance, materias genéricas o específicas. Esto supone que no es posible definir atributos específicos para cada una de las instancias de la clase materia que forman un sistema de clasificación ni, por tanto, asimilarlas a las clases de una ontología.

Otra diferencia importante reside en el hecho de que las ontologías modelan entidades, cosas, mientras que los lenguajes documentales trabajan con el nivel léxico en algunos casos y con el nivel conceptual en otros. Esto se materializa en la capacidad de las ontologías para modelar clases de entidades (con sus subclases correspondientes) e instancias (ejemplos concretos de uno de esos tipos de entidad) de forma diferenciada, mientras que en el ámbito de los lenguajes documentales esto no es posible.

"Una ontología sobre una universidad debería responder preguntas como: ¿Es el decano de una facultad responsable de las infraestructura de redes? o ¿Dónde debo presentar mis impresos de matriculación?"

3.2. Relaciones connotativas

Estas relaciones se sitúan en realidad entre la esfera semántica y la estructural en los lenguajes documentales. Se pueden dar entre los términos que componen los diferentes útiles de representación del conocimiento, que aquí estamos analizando. En el caso de los sistemas de clasificación, la connotación se da de forma particular en aquellos que organizan el conocimiento en base a disciplinas, es decir, por campos del saber, no por materias; con lo que una materia puede figurar en varios lugares de una misma clase y, a su vez, en diferentes clases del esquema clasificatorio. Son ejemplos de esta forma de organización la *Dewey Decimal Classification* (*DDC*) y la *Clasificación Decimal Universal* (*CDU*), siendo típica la organización por materias en la *Library of Congress Classification* (*LCC*).

Las relaciones connotativas derivadas de esta característica de algunos sistemas de clasificación, es decir, la posible existencia de un mismo término referido, no obstante, a distintas disciplinas o aspectos, se resuelve en los sistemas de clasificación mediante su representación con una numeración específica y unívoca para cada uno de sus potenciales significados. En los tesauros hablaríamos en este caso de polijerarquía, que se puede resolver mediante notas aclaratorias. Una ontología, por su parte, le puede indicar a una aplicación que dos términos polisémicos no tienen ninguna relación y pertenecen a distintas disciplinas.

Por lo que se refiere a la estructura, la que caracteriza a los tesauros es tradicionalmente arborescente, representada mediante árboles jerárquicos. Las ontologías son flexibles y multidimensionales, en contraste con la estructura jerárquica. Por su parte, los sistemas de clasificación por facetas no tienen la misma estructura arborescente de los modelos tradicionales, y además tienen en cuenta aspectos (facetas) que no son consideradas por aquellos. Por ejemplo, el sistema de clasificación de **Bliss** contempla las disciplinas desde los puntos de vista filosófico, teórico, histórico y prác-

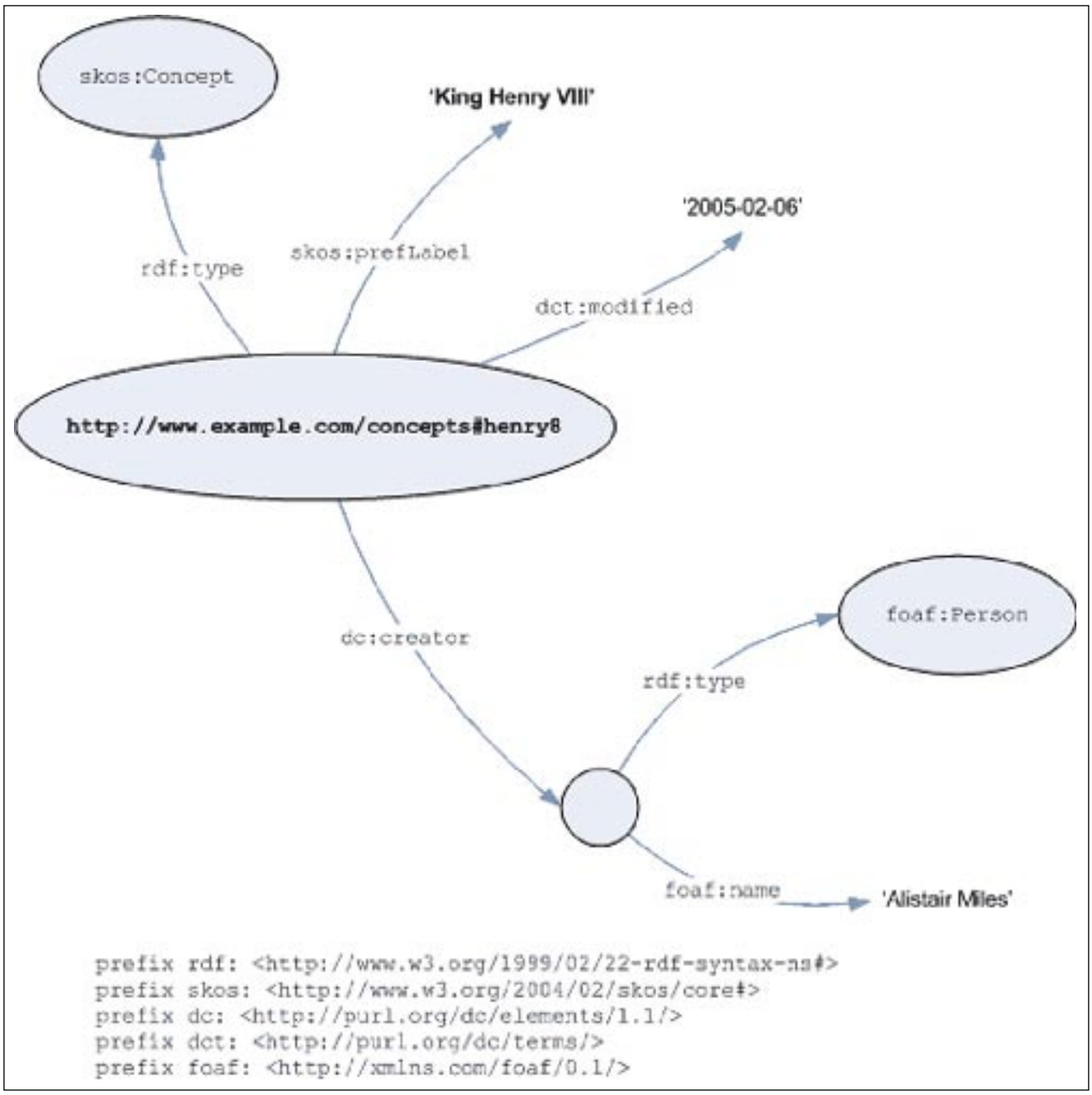


Figura 4. Capa de indirección en los lenguajes documentales. Fuente: SKOS Core Guide

tico; es decir, base teórica de la disciplina, la disciplina en sí misma, su historia y sus aplicaciones, y sustituye la ordenación jerárquica por una gradación de lo simple a lo complejo en las disciplinas y de lo general a lo particular entre los distintos conceptos de una materia.

La *Colon Classification* tampoco es un sistema de clasificación jerárquico como los tradicionales, ya que no contempla un esquema de conocimientos que se va organizando jerárquicamente, sino una "realidad básica", como afirmaba **Ranganathan**, que se descompone en facetas. Esto es, el sistema consiste en clasificar las materias no sólo por la relación jerárquica de género a especie, sino también por los vínculos existentes entre una cosa y sus partes, las materias que la componen, sus propiedades, los tratamientos a los que pueden se sometidos, el espacio y el tiempo. De esta forma, todos los conceptos son susceptibles de caracterizarse por una serie de aspectos o características comunes.

Estos mismos principios son los que inspiran a los tesauros facetados. Algunas de las facetas en las que suelen organizar el conocimiento son: fenómeno, procesos, materiales, propiedades, disciplina, etc. Desde el punto de vista estructural, su utilización proporciona seguramente el punto de conexión más viable entre un lenguaje documental y una ontología que, como veremos en otro punto, se basa en una concepción muy flexible de los tipos de relaciones posibles entre los diversos elementos del sistema.

3.3. Relaciones jerárquicas entre clases

Tienen un papel fundamental a la hora de definir la estructura de los lenguajes documentales. Utilizamos el término relaciones jerárquicas aunque quizá no describe completamente el caso, ya que las que se establecen entre una clase y una subclase en un sistema de clasificación, o entre los términos específicos y genéricos en un tesauro, no son exactamente del mismo tipo que las existentes entre las clases y las subclases de una ontología.

Las clases y las subclases de una ontología se relacionan entre sí mediante un mecanismo de *subsunción*, lo que implica que dada una clase *C* con una subclase C_1 si *m* es un miembro de C_1 también lo es de *C*. En otras palabras, si la clase "*bebida"* tiene una sub-clase "*vino"*, podemos decir que "*vino"* es una *"bebida"*. Esto implica que todos los atributos que son propios de *C* lo son también de C_1, aunque no se cumpla a la inversa, lo que se define comúnmente como un mecanismo de herencia (**Corcho**; **Fernández**; **Gómez**, 2003).

Podemos analizar el ámbito de los lenguajes documentales en torno a dos enfoques, el que es propio de los tesauros y el de las clasificaciones. En el primer caso, las relaciones jerárquicas pueden ser *grosso modo* de tipo partitivo o de generalidad/especificidad. En el primero de los casos se establece que un "*término específico"* lo es de otro porque constituye una parte de aquél, como en el caso de: *Italia TE Roma*. En el segundo caso "el término genérico se define como aquel descriptor que designa una noción que engloba a otras nociones más específicas representadas por los términos específicos" (**Gil**, 2004).

La norma *ISO 2788:1986* para la construcción de tesauros monolingües es muy ambigua al respecto de las relaciones jerárquicas, de forma que el significado exacto que ofrece depende del criterio de los expertos que crean cada tesauro en concreto. Esto ha llevado a la definición de relaciones mejor perfiladas por parte de los profesionales pero que habitan fuera del estándar. De esta manera se definen relaciones "*broader term generic"*, que recogen los términos genéricos en general, relaciones "*broader term instantive"*, que tienen en cuenta las relaciones de instancia entre una clase y sus miembros, o "*broader term partitive"*, que recogen las de una entidad y sus partes.

La tabla 1 muestra un extracto del tesauro *Agrovoc* y la reformulación del mismo en forma de ontología que hace **Soergel** (2004). De esta tabla se deduce fácilmente que la relación existente entre "milk" y "cow milk" no es en absoluto igual a la que existe entre "milk" y "milk fat". Esto se debe a que el tipo exacto de relación cubierto por los términos específicos se halla implícito. Esto es así porque su explicitación no es necesaria para un indizador humano. Cuando trabajamos en el ámbito de la indización y recuperación manual, un experto es perfectamente capaz de apreciar de forma directa los matices del tipo de relación y juzgar su idoneidad para cada situación. Sin embargo, no es en modo alguno irrelevante para una máquina que quisiera utilizar provechosamente la relación existente entre ambos conceptos. Y éste es precisamente el enfoque que se intenta lograr en la web semántica.

Agrovoc		Hypothetical ontology	
milk		**milk**	
NT	cow milk	*includesSpecific-*	cow milk
NT	milk fat	*containsSubstance-*	milk fat
cow		**cow**	
NT	cow milk	*hasComponent-*	cow milk
cheddar cheese		**cheddar cheese**	
BT	cow milk	*madeFrom-*	cow milk

Tabla 1. Relaciones jerárquicas. Comparación entre el tesauro Agrovoc y una hipotética ontología. Fuente: Soergel 2004

En el ámbito de los sistemas de clasificación ocurre algo similar. Tomemos por ejemplo la *CDU*; podemos encontrar cómo se produce con cierta frecuencia la relación entre materias y sus divisiones en torno al criterio de pertenencia a una clase, como en "616.91 *Enfermedades infecciosas febriles agudas"* que se divide en "616.912 *Variola"*, "616.913 *Varicela"*, "616.914 *Sarampión",* etc. Como se puede observar, "*Varicela"* es una instancia de la clase "*Enfermedades infecciosas febriles agudas"* y por tanto un miembro concreto de un grupo de enfermedades con una serie de características en común. No es un tipo de enfermedad, sino una enfermedad específica, por lo que en ningún caso estaríamos hablando de la relación de subsunción existente en el ámbito de las ontologías entre una clase y su subclase, sino de una relación clase/instancia bien distinta.

"La capacidad de precisión de las ontologías es fundamental para el procesamiento automático del conocimiento"

Otro fenómeno común es el de las relaciones "parte de" que se establecen entre muchas materias y sus divisiones, como entre "004.42 *Programación de ordenadores. Programas de ordenador"* y "004.421 *Algo-*

ritmos para la construcción del programa", "004.422 *Componentes de los programas de ordenador"*. Son tan sólo algunos ejemplos pero, en cualquier caso, son representativos de las diferencias existentes entre la estructura jerárquica típica de un lenguaje documental y la de una ontología. Esta capacidad de precisión de que hacen gala las ontologías es fundamental para el procesamiento automático del conocimiento.

3.4. Otros tipos de relaciones

Si observamos un tesauro tradicional, podemos comprobar cómo las relaciones asociativas se pueden reformular mediante múltiples tipos de relaciones más expresivas y formalmente correctas.

Eric Thesaurus	Hypothetical ontology
reading instruction	**reading instruction**
BT instruction	*isa-* instruction
RT reading	*hasDomain-* reading
RT learning standards	*governedBy-* learning standards
reading ability	**reading ability**
BT ability	*isa-* ability
RT reading	*hasDomain-* reading
RT perception	*supportedBy-* perception

Tabla 2. Relaciones asociativas. Comparación entre el tesauro Eric y una hipotética ontología. Fuente: Soergel 2004

Si consideramos los términos relacionados de "*reading ability"* podemos observar cómo la relación asociativa típica de los tesauros es poco explícita y se podría concretar en otras mucho más específicas y significativas, como "*hasDomain"* o "*supportedBy"*. En realidad los dos tipos de relación propuestos en la ontología no tienen nada que ver entre sí, son de hecho tan distintos como para hacer poco justificable que se agrupen bajo un mismo tipo genérico de relación, la asociativa.

Los tipos de relaciones que se pueden establecer entre las materias de un sistema de clasificación, especialmente en los sistemas de tipo mixto como la *CDU,* son bastante más expresivos, lo que nos dota de cierta capacidad de combinación y adaptación a las necesidades reales de la clasificación de documentos, aunque siguen siendo limitados, ya que existe un número finito de ellos, que no afecta a la definición de las relaciones entre conceptos en una ontología.

4. Conclusiones

Los lenguajes documentales no pueden utilizarse de la misma forma que las ontologías para tareas de formalización del conocimiento, ni siquiera desde la perspectiva de una funcionalidad reducida. A día de hoy no pueden emplearse como una base de conocimiento, porque no han sido construidos de forma que este conocimiento resulte explícito para su tratamiento automatizado.

Desde el punto de vista de la reutilización de los lenguajes documentales hemos de decir que las diferencias entre éstos y las ontologías complican la posibilidad de trasladar el conocimiento existente en un tesauro o un sistema de clasificación a una ontología. Reducir los tipos de relaciones que se establecen en los lenguajes controlados tradicionales haría de la ontología resultante un sistema inoperante para las tareas de razonamiento deductivo. Explicitar los tipos de relaciones es una tarea que, como mucho, se puede semi-automatizar, pero que en cualquier caso tendría un coste elevado en términos de tiempo y esfuerzo.

"Las diferencias entre lenguajes documentales y ontologías complican el traslado del conocimiento existente en un tesauro o un sistema de clasificación a una ontología"

Por otra parte, los lenguajes documentales pueden servir como base para el desarrollo de ontologías, como se ha hecho en varios de los artículos mencionados, pero tienen un papel limitado en este sentido. Trazando un símil con el proceso de creación de un tesauro, podemos comparar la eficacia que tendrían los lenguajes documentales para la creación de ontologías con la utilidad que tienen los glosarios o listados de términos utilizados durante la fase de extracción de terminología. Los lenguajes documentales existentes proporcionan un acercamiento a los conceptos típicos de un dominio, pero tendrían un papel mucho más limitado a la hora de construir la estructura, y sobre todo para la formalización de las clases y el establecimiento de reglas y axiomas formales.

Parece más factible que los lenguajes documentales puedan adaptarse a los requerimientos de una ontología para la recuperación de información. Se han llevado a cabo numerosos estudios sobre la utilidad de este tipo de ontologías (**Bhogal**, et. al., 2007; **Paralic**; **Kostial**, 2003; **Vallet**, et. al., 2005) aunque los resultados no son concluyentes.

En cualquier caso nosotros nos hemos centrado en una visión sobre las ontologías para su utilización en el ámbito de la web semántica, lo que implica capacidad para expresar formalmente el significado de las cosas y la posibilidad de reutilización del conocimiento en

entornos diferentes al original. En este contexto los lenguajes documentales están muy alejados de los requerimientos formales de una ontología señalados por el *W3C*, como se ha expuesto con anterioridad. Por estos motivos creemos que la reutilización de los lenguajes documentales dentro del ámbito de la web semántica requerirá bastantes esfuerzos por parte de la comunidad de profesionales y expertos en lenguajes documentales.

Notas

1. En el original: "*the semantic web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation*".

2. Véase: **McBride, Brian** (2004) para una explicación en profundidad de éstos y otros aspectos básicos de rdf.

3. En el original: "*a formal explicit description of concepts in a domain of discourse*".

4. En el original: "*an abstract, simplified view of the world that we wish to represent for some purpose". "A conceptualization is an abstract, simplified view of the world that we wish to represent for some purpose*".

5. Algunos sitios web de referencia como *Amazon* o *Yahoo* hacen uso de ontologías como forma de gestión del conocimiento, una práctica que parece estar extendiéndose paulatinamente.

Bibliografía

Bechhofer, Sean; **Van Harmelen, Frank**; **Hendler, Jim**; **Horrocks, Ian**; **McGuinness, Deborah L.**; **Patel-Schneider, Peter F.**; **Stein, Lynn Andrea**; **Olin, Franklin W.** *OWL Web ontology language*. *Recomendación W3C*, febrero 2004. Consultado en: 12-02-06.
http://www.w3.org/TR/owl-ref/

Berners-Lee, Tim. *Transcripción de la conferencia en la conmemoración del 35 aniversario del Instituto Tecnológico de Massachusets el 14 de abril de 1999*.
http://www.w3.org/1999/04/13-tbl.html

Bhogal, Jagdev; **MacFarlane, Andrew**; **Smith, Peter**. "A review of ontology based query expansion". En: *Information processing & management*, 2007, julio, v. 43, n. 4, pp. 866-886.

Corcho, Óscar; **Fernández-López, Mariano**; **Gómez-Pérez, Asunción**. "Methodologies, tools and languages for building ontologies. Where is their meeting point?". En: *Data and knowledge engineering*, 2003, n. 46, pp. 41-64.

García-Jiménez, Antonio. "Instrumentos de representación del conocimiento: tesauros versus ontologías". En: *Anales de documentación*, 2004, n. 7, pp. 79-95.

Genesereth, Michael J.; **Nilsson, Nils J.** *Logical foundations of artificial intelligence*. San Francisco: Morgan Kaufmann, 1987.

Gil-Urdiciain, Blanca. *Manual de lenguajes documentales*. Gijón: TREA, 2004.

Gruber, Thomas R. "Towards principles for the design of ontologies used for knowledge sharing". En: **Guarino, N.**; **Poli, R.** (eds.). *Formal ontology in conceptual analysis and knowledge representation*. Deventer: Kluwer Academic Publishers, 1993.

Guzmán-Luna, Jaime; **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". En: *Revista interamericana de bibliotecología*, 2006, julio-diciembre, v. 29, n. 2, p.79-94.

Heflin, Jeff. *OWL Web ontology language use cases and requirements*. *Recomendación W3C*, febrero 2004.
http://www.w3.org/TR/webont-req/

Maniez, Jacques. *Los lenguajes documentales y de clasificación: concepción, construcción y utilización en los sistemas documentales*. Madrid: Fundación Germán Sánchez Ruipérez, 1992, p. 22.

McBride, Brian. *Rdf primer*. *Recomendación W3C*, febrero 2004.
http://www.w3.org/TR/rdf-primer/

McGuinness, Deborah L.; **Van Harmelen, Frank** (eds.). *OWL Web ontology language overview*. *Recomendación W3C*, febrero 2004.
http://www.w3.org/TR/owl-features/

Miles, Alistair. *Modelling thesauri for the semantic web*, *W3C*, 2001.
http://www.w3c.rl.ac.uk/SWAD/thesaurus/tif/deliv81/final.html#sec-owl

Miles, Alistair. *SKOS core guide*. *W3C Working Draft*, mayo 2005.
http://www.w3.org/TR/swbp-skos-core-guide/

Miller, Eric. *Semantic web activity statement.* W3C, noviembre 2005.
http://www.w3.org/2001/sw/Activity

Noy, Natalya F.; **McGuinness, Deborah L.** *Ontology development 101: a guide to creating your first ontology*. Informe técnico. Universidad de Stanford, 2000.
http://protege.stanford.edu/ontology101-noy-mcguinness.html

Paralic, Jan; **Kostial, Ivan**. "Ontology-based information retrieval" En: *Proceedings of the 14th International conference on information and intelligent systems, IIS 2003*, pp. 23-28.

Pérez-Agüera, José-Ramón. *Generación automática de tesauros documentales*. Trabajo para la obtención del diploma de estudios avanzados. Dirigido por **Lourdes Araujo**. Universidad Complutense de Madrid, septiembre de 2005, pp. 107-112.

Soergel, Dagobert; **Liang, Anita**; **Lauser, Boris**; **Fisseha, Frehiwot**; **Keizer, Johannes**; **Katz, Stephen**. "Reengineering thesauri for new applications: the Agrovoc example". En: *Journal of digital information*, 2004, v. 4, n 4.

Vallet, David; **Fernández, Miriam**; **Castells, Pablo**. "An ontology-based information retrieval model". En: *Proceedings of the Second European semantic web conference*, *ESWC 2005*, 2005, pp. 455-470.

Rodrigo Sánchez-Jiménez; Blanca Gil-Urdiciain,
Facultad de Ciencias de la Información, Universidad Complutense de Madrid.
rsanchezj@ccinf.ucm.es
blanca@caelo.eubd.ucm.es

De repente, ¿todos hablamos de ontologías?

Por **Sonia Sánchez-Cuadrado, Jorge Morato-Lara, Vicente Palacios-Madrid, Juan Llorens-Morillo** y **José Antonio Moreiro-González**

Sonia Sánchez-Cuadrado *es licenciada en Filología por la Universidad Complutense de Madrid y doctora en Documentación por la Universidad Carlos III. Desde 2002 es profesora ayudante en el Departamento de Informática de dicha Universidad con docencia en la licenciatura de documentación e informática sobre temas relacionados con modelado e ingeniería de la información. Actualmente se halla involucrada en distintos proyectos para la creación automática de ontologías.*

Jorge Morato-Lara *es licenciado en CC. Biológicas por la Universidad de Alcalá de Henares. A partir de 1991 empezó a trabajar en distintas empresas relacionadas con la documentación. Obtuvo el título de doctor en documentación por la Universidad Carlos III en 1999. Desde ese mismo año ha impartido clases en las licenciaturas de Documentación, ADE e Informática en asignaturas relacionadas con la ingeniería de la información, recuperación de información e ingeniería del software.*

Vicente Palacios-Madrid *es ingeniero superior por la Universidad Carlos III de Madrid. Es profesor de ingeniería del software e ingeniería de la información y miembro del servicio informático de la misma unidad. Actualmente se encuentra realizando la tesis en Documentación sobre interoperabilidad semántica de ontologías.*

Resumen: *Los artículos sobre ontologías llevan casi veinte años de gran actualidad en la literatura profesional, como se puede comprobar haciendo búsquedas en bases de datos. A pesar de esto, sigue sin haber consenso sobre el significado de este concepto. Probablemente la representación gráfica conocida como el espectro de las ontologías ha generado cierta confusión en algunos lectores. Por ello tratamos de argumentar aquí el origen de esta confusión, que puede deberse a la mezcla de varios tipos de sistemas de organización del conocimiento, cuya distinta finalidad llevó a definirlos de diferente forma. De hecho, construir una ontología formal para un sistema no siempre supone una mejora y en muchas ocasiones es suficiente con una representación menos compleja.*

Juan Llorens-Morillo *es ingeniero industrial y catedrático de Informática en la Universidad Carlos III de Madrid. Sus principales áreas de investigación se centran en la representación y procesamiento de la información para la reutilización del software.*

José-Antonio Moreiro-González *es catedrático de Documentación y Decano de la Facultad de Humanidades, Comunicación y Documentación de la Universidad Carlos III de Madrid. Ha escrito diferentes monografías y artículos sobre diversos aspectos de Biblioteconomía y Documentación.*

Palabras clave: *Ontologías, Tesauros, Sistemas de organización del conocimiento, Espectro de las ontologías, Tipos de ontologías*

Title: **And suddenly, everybody is talking about ontologies?**

Abstract: *The number of papers written on ontologies has increased considerably over the last twenty years. This trend can be easily observed by searching words like "ontology" or "thesaurus" in databases. Despite this fact, there hasn't been a consensus about the significance of this concept. The graphical representation known as ontology spectrum must have generated a lot of confusion amongst readers. In this paper we argue that this confusion is due to the mix of the various types of knowledge organization systems with distinct objectives in the same graphical representation. Thus, constructing a formal ontology for a system does not always presume an improvement, frequently it is adequate with less complex representations.*

Keywords: *Ontologies, Thesauri, Knowledge organization systems, Ontology spectrum, Types of ontologies*

Sánchez-Cuadrado, Sonia; **Morato-Lara, Jorge**; **Palacios-Madrid, Vicente**; **Llorens-Morillo, Juan**; **Moreiro-González, José Antonio**. "De repente, ¿todos hablamos de ontologías?". En: *El profesional de la información,* 2007, noviembre-diciembre, v. 16, n. 6, pp. 562-568.

DOI: 10.3145/epi.2007.nov.03

Artículo recibido el 13-07-07
Aceptación definitiva: 26-09-07

Es de sobras conocido que las ontologías son un concepto bastante antiguo entre los mayores predicados de la Filosofía. Sin embargo, últimamente, el término "ontología" ha cobrado gran relevancia, siendo posiblemente uno de los que más definiciones y menciones acumula desde principios de los años noventa y, en especial, por parte de la comunidad científica de áreas de conocimiento ajenas a la Filosofía.

¿Qué son las ontologías?

Entre las definiciones más extendidas se encuentran las de **Gruber** y **Guarino**. Para **Gruber** (1993), las ontologías son una especificación de una conceptualización. Mientras, **Guarino** (1998) definió la ontología como un producto de ingeniería consistente en un vocabulario específico usado para describir una realidad más un conjunto de asunciones relacionadas con el significado del vocabulario.

Por otra parte, **Lassila** y **McGuinness** (2001) hablan del espectro de las ontologías, *ontology spectrum*, que no es más que un espacio en el cual se presentan diferentes formas de representar el conocimiento. El marco en el que se desarrolla la comparación toma como elemento principal la riqueza semántica, caracterizando los extremos por semántica débil y semántica fuerte. Aquello que se encuentra próximo al extremo débil representa una semántica simplificada, mientras que lo que se encuentra en el extremo fuerte representa una semántica más compleja. Posteriormente se han realizado revisiones como la de **Daconta** et al. (2003: 157).

En otras palabras, el espectro de las ontologías (Ilustración 1), considera como ontología a toda organización cognitiva que oscile desde la noción más simple de las taxonomías, pasando por los tesauros y modelos conceptuales hasta llegar a las teorías lógicas que representan la noción más compleja. Desde este punto de vista, una ontología define conceptos (significados) usados para describir y representar un área de conocimiento.

¿Por qué la comunidad científica no se pone de acuerdo en qué es una ontología y qué no?

En este punto los expertos se preguntan y discuten si las taxonomías, los tesauros o los mapas conceptuales se consideran también ontologías. En cualquier evento científico donde se hable de ontologías, son comunes dos frases: "[…] sí claro, eso también es una ontología […]" y/o "[…] no, es que eso no es una ontología […]". Tal es así, que aunque **Studer** et al. (1998) incorporaron en la definición los términos conceptualización explícita, formal y compartida (esto último como conocimiento consensuado por un grupo o comunidad), aún hoy paradójicamente el término "ontología" sigue sin tener para muchos una significado consensuado.

Probablemente la comunidad científica discrepa a la hora de determinar qué es una ontología y qué no lo es debido a que las definiciones propuestas tienen una vertiente específica (cuando se definen las ontologías formales), pero también otra genérica (referida a las ligeras).

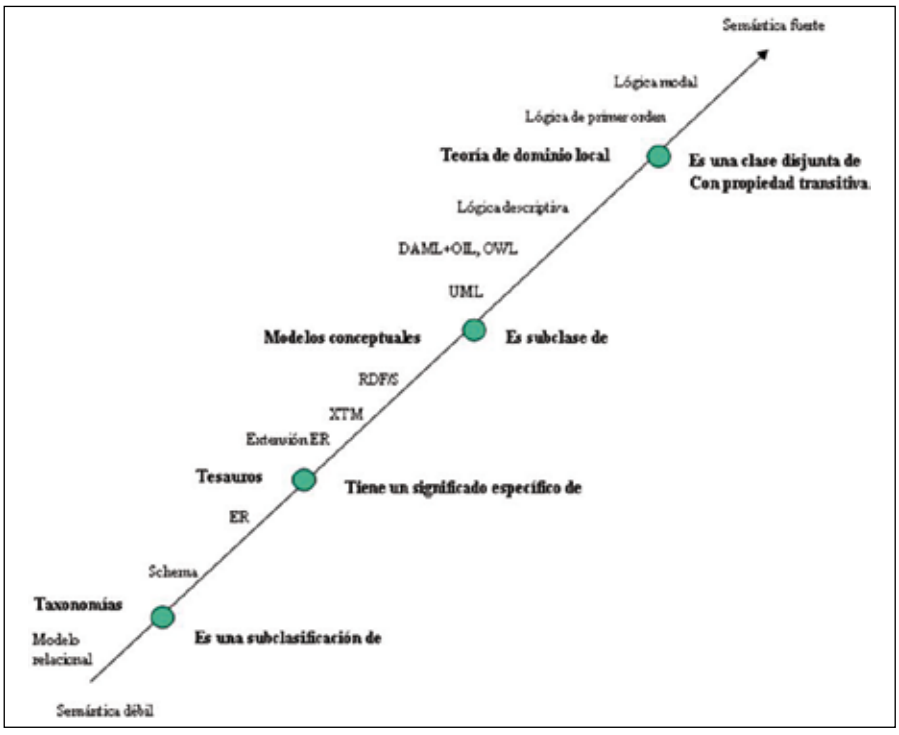


Ilustración 1: Espectro de las ontologías

En consecuencia, cada disciplina (Documentación, Ingeniería lingüística, Ingeniería de software, Inteligencia artificial, etc.) ha adaptado la definición de las ontologías a sistemas que venían desarrollando antes (tesauros, redes semánticas, modelos conceptuales, etc.). Y así, el espectro de las ontologías alberga un considerable número de conceptos asociados a áreas de conocimiento que hasta el momento rara vez habían sido denominados así.

Por otra parte, es posible que la ilustración del espectro de las ontologías incite a cierta confusión al dibujar en la misma línea modelos de conocimiento y lenguajes para representar el conocimiento.

¿Cuántos tipos de ontologías hay?

Cuando la comunidad científica diferencia entre tipos de ontologías suele clasificarlas siguiendo diferentes criterios: según el contenido (**Mizoguchi** et al., 1995); según el volumen y el tipo de estructura, pero también por la conceptualización específica del conocimiento (**Heijst** et al., 1997); según su grado de dependencia -ontologías de alto nivel, de dominios genéricas, de dominio, de aplicación, y de tareas- (**Guarino**, 1998); o por la riqueza semántica de su estructura interna (**Lassila** y **McGuinness**, 2001).

> "Se considera ontología toda organización cognitiva"

Siguiendo a otros autores, habría que añadir además las denominadas ontologías lingüísticas (**Gómez-Pérez** et al., 2004) que consideran las palabras como unidades gramaticales y están pensadas para describir construcciones semánticas, más que para proponer modelos específicos de un dominio. La finalidad de las ontologías lingüísticas, por ejemplo *WordNet* (**Miller** et al., 1990) o *Sensus* (**Swartout** et al., 1997), ha sido conformar bases de datos léxicas, aplicándolas a las máquinas de traducción y a la generación de lenguaje natural. Sin embargo, no toda la comunidad científica coincide en otorgar a recursos como *WordNet* la categoría de ontología, a pesar de que a partir suyo se hayan generado ontologías como *Sensus* o *webKB*.

¿Estamos frente a una burbuja ontológica?

La respuesta a esta pregunta es afirmativa, aunque sólo sea parcialmente. Sin duda es un término de moda, del que se han escrito infinidad de documentos que vinculan las ontologías con múltiples áreas de conocimiento, provocando incluso cierta descompensación entre la producción literaria y el número de ontologías existentes.

Una revisión de diferentes bases de datos demuestra esta casuística. Se seleccionan como ejemplo los términos "ontología" y "tesauro", si bien puede comprobarse que se produce el mismo comportamiento con otros términos referidos a los sistemas de organización del conocimiento. Para ello se realizó una búsqueda con los términos "*ontology*" y "*thesaurus*" (junto con sus variaciones gramaticales) para los años 1985, 1995, 2000 y 2003 en las bases de datos *COS Scholar universe: Social science* e *Ieee* (esta última especializada en artículos de ingeniería a texto completo). Los resultados muestran una ligera tendencia incremental en el uso del término "tesauro" y una fuerte tendencia exponencial en el de "ontología" (tabla 1).

	COS Scholar social		Ieee	
	thesaurus	ontología	thesaurus	ontología
1985	27	159	0	0
1995	44	263	7	11
2000	48	326	11	50
2003	46	417	14	205

Tabla 1. Utilización de los términos ontología y tesauros en las bases de datos COS de CSA e Ieee full-text.

Un análisis de la base de datos de referencia *Research index* muestra un fenómeno análogo.

Si bien la elaboración de una ontología presenta una serie de etapas comunes con la propia de los tesauros, resulta sugestivo comparar cuándo se utiliza un término y no el otro, y cuándo se utilizan de forma conjunta. Dicha comparación se ha realizado con el buscador web de documentación científica *Google Scholar*, y los resultados se pueden ver en la siguiente tabla (tabla 2).

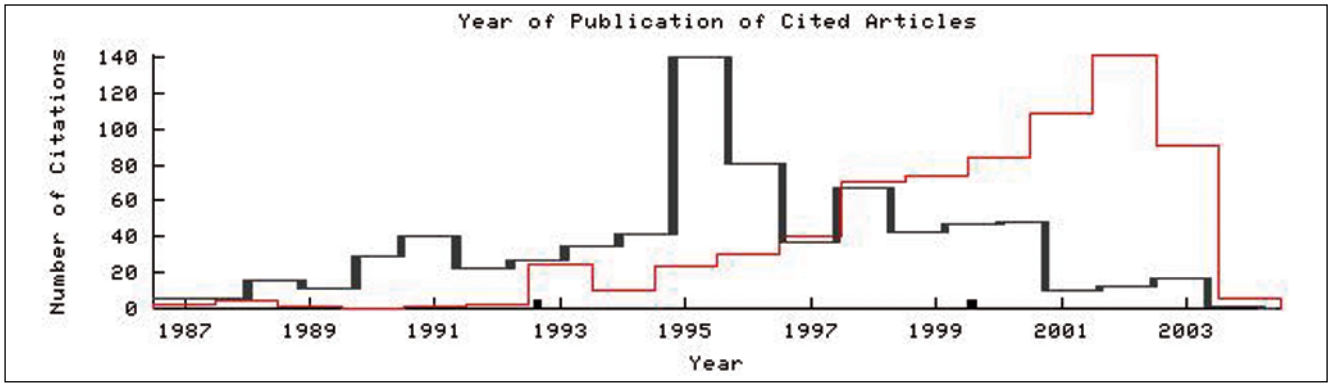


Ilustración 2: Comparación del uso del término "ontologías" (trazo fino) y "tesauros" (trazo grueso) durante el período 1987-2003 según el análisis de citas de ResearchIndex

	Ontologías sin tesauros	Tesauros sin ontologías	Ontologías y tesauros
1985	614	371	4
1995	2.480	1.780	48
2000	6.580	4.180	265
2003	12.800	3.780	608
2006	15.000	3.010	657

Tabla 2. Uso de los términos "ontologías" y "tesauros" en Google Scholar

En esta tabla se observa el uso creciente del término "ontología", y la tendencia inversa para "tesauro", y cómo algunas instancias del término "ontologías" están hablando también de tesauros.

Por otra parte, no es extraño que se escriba tanto sobre las ontologías, no sólo por la controversia del concepto, sino por tratarse de un tema complejo sobre el que existen infinidad de cuestiones pendientes de resolver.

¿Por qué han cobrado tanta importancia las ontologías?

La notable presencia alcanzada por las ontologías puede deberse a su consideración como recurso fundamental para la web semántica (**Tim Berners-Lee**, 1998). Posiblemente estemos ante otro término de moda, cuya función sea dar respuesta a la necesidad de describir la semántica de un dominio en el entorno web de modo que pueda ser interpretada por las aplicaciones informáticas de forma inteligible para los usuarios.

Asimismo, parte de la relevancia alcanzada puede atribuirse a que durante décadas se ha estado investigando en lo que algunos ahora denominan ontologías y previamente eran tipos de sistemas de organización del conocimiento o *KOS* (*knowledge organization systems*) (**Zeng**; **Chan**, 2004).

¿Por qué hay ontologías en todos los ámbitos?

Resulta evidente la necesidad de estructurar semánticamente los conceptos en determinado sistema de organización del conocimiento. El fundamento de las ontologías es representar el conocimiento teniendo como punto de partida la organización taxonómica. Una taxonomía puede ser descrita como un modo de clasificar o categorizar un conjunto de conceptos de forma jerárquica, siguiendo básicamente una estructura en forma de árbol que establece entre los conceptos una relación de generalización-especialización (donde las relaciones son "subclasificación-de" o "subclase-de") y dentro, por tanto, de una semántica simple si seguimos 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 las taxonomías a las ontologías más completas y formalizadas.

La semejanza de estos sistemas a lo que en inteligencia artificial y la web semántica se denominaba "ontologías" resultaba evidente, siendo importado el término por otras áreas de conocimiento. Por tanto, uno de los motivos que justifica el uso de ontologías en diferentes disciplinas es precisamente que ya existía una forma de representar el conocimiento de un dominio específico, pero recibiendo hasta ahora otra denominación. De otra parte, ciertamente las ontologías suponen un recurso útil para multitud de sistemas de los que se espera que cuenten con relaciones entre sus elementos y cierto grado de inferencia o razonamiento.

Además, hay que tener en cuenta que la ausencia de un estándar que clarifique cuál es el mínimo conjunto de elementos que una ontología debe contener o restringir y la finalidad a la que se debe aplicar, convierten a este término en un "cajón de sastre".

Y entonces...

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. Para ello, desde diferentes disciplinas (Biblioteconomía y Documentación, Inteligencia artificial, Ingeniería del software, Lingüística, Ingeniería ontológica, etc.) se ha propuesto una serie de modelos para su representación (**Llorens** et al., 2004: 221-253).

El grado de representación semántica y la finalidad que se requiera plasmar condicionará los modelos y lenguajes a la hora de construir un sistema de organización del conocimiento, entendiendo como tales aquellos que están englobados en el espectro de las ontologías.

Las ontologías como sistemas de organización del conocimiento tratan de representar tanto información genérica como información concreta. 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. En muchos casos, los lenguajes para los modelos de representación del conocimiento representan un paradigma completo y un lenguaje de soporte.

Dentro de la web semántica, las ontologías constituyen una pieza clave para el modelado de conocimiento (**Antoniou**; **Harmelen**, 2004: 10-18). Las principales

adaptaciones que han sufrido éstas al entrar en contacto con internet han sido:

– la habilitación de elementos para mejorar la interoperabilidad y reutilización en un entorno heterogéneo (como es internet);

– la expresión mediante lenguajes web (como xml o basados en éste).

El potencial de las ontologías dentro de la web semántica está determinado por la capacidad desarrollada a la hora de favorecer la interoperabilidad y la capacidad de reutilización. Estas propiedades se fundamentan en:

– La adopción de lenguajes comunes y compatibles (con un lenguaje de sintaxis, xml, y la expresión

Disciplina	Objetivos	Elementos máximos	Ejemplos de representación del conocimiento	Carga semántica
Documentación	Representar el conocimiento de un dominio	términos	Listados de términos	- complejo
	Clasificar objetos de información	*conceptos*	Taxonomías	
	Recuperar información	sinonimias	Tesauros	
		taxonomías	Topic Maps	
		relaciones asociativas		+ complejo
Ingeniería del software	Modelado de bases de datos	conceptos	Modelos *ER*	- complejo
	Modelado de aplicaciones	taxonomías	Diagrama *ER* extendido	
	Ayudas gráficas para comunicación entre clientes, analistas y desarrolladores	relaciones asociativas	Diagramas de clases de *UML*	+ complejo
		funciones		
		restricciones básicas		
Inteligencia artificial	Imitar la mente humana	conceptos	Redes semánticas	- complejo
	Almacenar conocimiento común	sinonimias	*OWL DL* (lógica descriptiva)	
	Con mecanismos para realizar inferencias	taxonomías	*SCL* (lógica de primer órden)	
	Con mecanismos de aprendizaje	relaciones asociativas		+ complejo
	Con capacidad de operar en sistemas informáticos	funciones		
		restricciones formales		
		reglas de inferencia		

Tabla 3. El espectro de las ontologías desglosado según la finalidad de las disciplinas de Documentación, Ingeniería del software e Inteligencia artificial

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*).

Por lo tanto, y a modo de conclusión, el espectro de las ontologías sería más clarificador y completo añadiendo otros parámetros complementarios como la finalidad para la que se construye el recurso. Estos objetivos, en función del dominio de aplicación, se han esquematizado en las tablas 3 y 4.

Las tablas 3 y 4 presentan ejemplos de sistemas de organización del conocimiento propuestos desde diferentes disciplinas. Se observa como los elementos máximos que los caracterizan están condicionados por el área de conocimiento y el objetivo para el que habían sido propuestos.

Para la ingeniería ontológica, las ontologías pueden estar organizadas con distintas técnicas de modelado de conocimiento y ser implementadas en diversos lenguajes (**Uschold**; **Grüninger**, 1996). Sin embargo, los diferentes modelos y técnicas no permiten representar el mismo conocimiento y, por tanto, tampoco tienen el mismo grado de especificación semántica.

En resumen, una ontología puede tener mayor o menor grado de complejidad y puede ser visualizada mediante diversos criterios gráficos. No obstante, los diferentes tipos de ontologías o *KOS* que se deban implementar estarán en función de los objetivos de aplicación (tablas 3 y 4). Bajo este principio las ontologías se pueden clasificar según su finalidad, e incluso pueden denominarse de distinta manera dependiendo de la disciplina. Así, no siempre resulta rentable el esfuerzo de realizar una ontología formalizada con axiomas y reglas de inferencia si la finalidad del recurso no lo justifica.

Bibliografía

Antoniou, G.; **Harmelen, F. van**. *A semantic web primer*. London: The MIT Press, 2004.

Berners-Lee, Tim. *Semantic web road map. Personal notes.* 1998. *http://www.w3.org/DesignIssues/Semantic.html*

Daconta, Michael C.; **Obrst, Leo J.**; **Smith, Kevin T.** *The semantic web. A guide to the future of XML, web services, and knowledge management*. Indianapolis: Wiley, 2003.

Gómez-Pérez, A.; **Fernández-López, M.**; **Corcho, O.** *Ontological engineering: with examples from the areas of knowledge management, e-commerce and the semantic web*. London: Springer, 2004. p. 403.

Gruber, T. R. "A translation approach to portable ontology specifications". En: *Knowledge acquisition*. Jun 1993, vol. 5, n. 2, pp. 199-220.

Guarino, N. "Formal ontology and information systems". En: *Proceedings of the 1st International conference on formal ontologies in information systems*, Trento, Italy, 6-8 Jun 1998, Amsterdam: IOS Press, 1998, pp. 3-15.

ISO/IEC 13250: 2000, *Information technology—SGML applications—topic maps*. Geneva: ISO, 2000. *http://www1.y12.doe.gov/capabilities/sgml/sc34/document/0129.pdf*

ISO/IEC 13250:2000, *Topic maps: information technology - Document description and markup languages*. **Michel Biezunski**, **Martin Bryan**, **Steven R. Newcomb** (eds.), 3 Dec 1999.

ISO-2788: 1986. *Guidelines for the establishment and development of monolingual thesauri*. International Organization for Standardization, Second edition 11-15 UDC 025.48. Geneva: ISO, 1986.

Disciplina	Objetivos	Elementos máximos	Ejemplos de representación del conocimiento	Carga semántica
Web semántica	Representar el conocimiento	conceptos	Vocabularios de metadatos	- complejo
	Clasificar objetos de información	sinonimias	rdf (*S*), *Topic maps*	
	Recuperar información	taxonomías	OWL Lite	
	Operar en sistemas heterogéneos con fines similares a la documentación, I.s. e I.a.	relaciones asociativas	OWL DL	+ complejo
	Formalización. Establecer Expresar de forma normalizada el conocimiento	funciones		
	Interoperabilidad	restricciones formales		
		reglas de inferencia		

Tabla 4. Las ontologías según la Web semántica

Lassila, Ora; **McGuinness, Deborah**. "The role of frame-based representation on the semantic web". KSL tech report number KSL-01-02. Jan 2001. *http://www.ksl.stanford.edu/people/dlm/etai/lassila-mcguinness-fbr-sw.html*

Llorens, Juan; **Morato, Jorge**; **Génova, Gonzalo**. "RSHP: An information representation model based on relationships". En: **Damiani, Ernesto**; **Jain, Lakhmi C.**; **Madravio, Mauro** (eds.). *Soft computing in software engineering*. Berlin: Springer, 2004. pp. 221-253. (Studies in fuzziness and soft computing series, vol. 159).

Miller, George; **Beckwith, Richard**; **Fellbaum, Christiane** (et al). "Introduction to WordNet: An on-line lexical database". En: *International journal of lexicography*, 1990, v. 3, n. 4, pp. 235-244.

Mizoguchi, R.; **Vanwelkenhuysen, J.**; **Ikeda, M.** *Task ontology for reuse of problem solving knowledge*. En: **Mars, N.** (ed.). *Towards very large knowledge bases: knowledge building and knowledge sharing (KBKS'95)*. University of Twente, Enschede, The Netherlands. Amsterdam: IOS Press, 1995, pp. 46-57.

Niso. 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*

Studer, R.; **Benjamins, V. R.**; **Fensel, D.** "Knowledge engineering: principles and methods". En: *Data & knowledge engineering*. Mar 1998, v. 25, n. 1-2, pp. 161-197.

Swartout, B.; **Patil, R.**; **Knight K.** (et al). "Toward distributed use of large-scale ontologies". En: *AAAI-97 Spring. Symposium series on ontological engineering*, 1997.

Uschold, M.; **Gruninger, M**. "Ontologies: principles, methods and applications". En: *Knowledge engineering review*, 1996, v. 11, n. 2, pp. 93-136.

Van Heijst, G.; **Schreiber, A. T.**; **Weilinga, B. J.** "Using explicit ontologies in KBS development". En: *International journal of human-computer studies*. 45:183-292. 1997.

Zeng, M. L.; **Chan, Lois M.** « Trends and issues in establishing interoperability among knowledge organization systems". En: *Journal of the American Society for Information Science and Technology*, 55, 5, 2004. pp. 377-395.

***Sonia Sánchez-Cuadrado**, **Jorge Morato-Lara**, **Vicente Palacios-Madrid**, **Juan Llorens-Morillo**, Departamento de Informática, Universidad Carlos III.*
***José-Antonio Moreiro-González**, Departamento de Documentación. Universidad Carlos III.*
ssanchec@ie.inf.uc3m.es
jorge@ie.inf.uc3m.es
palacios@di.uc3m.es
llorens@ie.inf.uc3m.es
jamore@bib.uc3m.es

Web semántica y ontologías en el procesamiento de la información documental

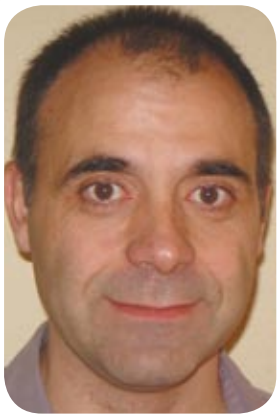
Por **Rafael Pedraza-Jiménez, Lluís Codina** y **Cristòfol Rovira**

Resumen: *La carencia de un modelo bien definido de representación de la información en la web ha traído consigo problemas de cara a diversos aspectos relacionados con su procesamiento. Para intentar solucionarlos, el W3C, organismo encargado de guiar la evolución de la web, ha propuesto su transformación hacia una nueva web denominada web semántica. En este trabajo se presentan las posibilidades que ofrece este nuevo escenario, así como las dificultades para su consecución, prestando especial atención a las ontologías, herramientas de representación del conocimiento fundamentales para la web semántica. Por último, se analiza el papel del profesional de la biblioteconomía y documentación en este nuevo entorno.*

Palabras clave: *Web semántica, Ontologías, Rdf, Owl, Sistemas de información.*

Rafael Pedraza-Jiménez *es miembro del grupo de investigación DigiDoc y profesor del Área de Biblioteconomía y Documentación de la Universidad Pompeu Fabra. Imparte docencia en las titulaciones de comunicación audiovisual y publicidad y relaciones públicas, así como en el máster online en documentación digital. Sus principales líneas de trabajo son las taxonomías y la generación semiautomática de ontologías, uno de los temas centrales de su tesis doctoral.*

Lluís Codina *es profesor titular de universidad. Imparte docencia en los estudios de periodismo y en la Facultad de Comunicación Audiovisual de la Universidad Pompeu Fabra de Barcelona. Es el investigador principal del Grupo de Investigación DigiDoc de la misma universidad. Participa en el máster interuniversitario UB/UPF en gestión de contenidos digitales, en el programa de doctorado del Departamento de Periodismo y de Comunicación Audiovisual y es co-director del máster online de documentación digital.*



Cristòfol Rovira *es profesor de la Universidad Pompeu Fabra en el Área de Biblioteconomía y Documentación. Imparte docencia en las titulaciones de publicidad y relaciones públicas y traducción e interpretación. Es coordinador del máster interuniversitario UB/UPF en gestión de contenidos digitales y director del máster online de documentación digital. Es investigador del grupo DigiDoc de la Universidad Pompeu Fabra y director del Laboratorio DigiDoc del mismo grupo.*

Title: Semantic web and ontologies in document information processing

Abstract*: The lack of a well defined model of information representation on the web has produced several problems related to processing information. In an effort to resolve these problems, the W3C has proposed the semantic web project. This new scenario offers both possibilities and difficulties for the future. Special attention is given to ontologies, fundamental tools for the representation of knowledge on the semantic web. Finally, the role of library and information professionals is considered in this new context.*

Keywords*: Semantic web, Ontologies, Rdf, Owl, Information systems.*

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

DOI: 10.3145/epi.2007.nov.04

1. Contexto

Hasta la primera mitad del siglo pasado la gestión de la información documental fue un dominio casi exclusivo de bibliotecarios, archiveros y documentalistas. Pero la introducción de los ordenadores en la segunda mitad del siglo XX, la continuada adaptación de los procesos de trabajo a las nuevas tecnologías y, principalmente, la creación de la web en los noventa supuso la incorporación de nuevas disciplinas (muy particularmente la teoría de la recuperación de información) a este entorno. La consecuencia inmediata ha sido la proliferación, desde entonces, de multitud de investiga-

Artículo recibido el 11-07-07
Aceptación definitiva: 18-10-07

ciones centradas en el desarrollo de tecnologías y métodos que permitan la organización y la gestión de la información documental.

No obstante, a pesar de los importantes avances aportados por las nuevas tecnologías, el usuario de la web aún carece de un sistema que permita procesar y acceder a la información documental contenida en sitios web de una manera fiable. El problema estriba en al menos tres aspectos: en primer lugar la web es un sistema descentralizado y heterogéneo completamente distinto de los escenarios para los que estaban más o menos bien preparadas las disciplinas clásicas vinculadas con la documentación y la recuperación de la información. En segundo lugar, lo que sucede en la www es una recuperación de información "con adversario" (*adversarial information retrieval*), otro aspecto nunca contemplado por la recuperación de información clásica. Por último, originalmente el método de marcado de la información, html, combina elementos de contenido con otros de presentación. Para un ser humano no hay ningún problema en interpretar el título de un documento a partir, por ejemplo, de su preeminencia, su formato y su lugar en la página, pero si el autor ha marcado el título con un elemento de formato (<*font*>) en lugar de uno semántico (<*h1*>), para un ordenador resultará imposible identificar el título.

"La web es un escenario heterogéneo completamente distinto de los que usaban las disciplinas clásicas vinculadas con la documentación"

Posteriores correcciones, como la firme recomendación del *W3C* de separar contenido de presentación mediante el uso de "nuevas" versiones de html y xhtml, así como el conjunto de especificaciones en torno al estándar xml se espera que, poco a poco, vayan alterando este panorama hacia una web donde el marcado de los documentos se realice de forma "semántica", es decir, utilizando etiquetas que expresen el significado de los elementos y no su formato (que queda a cargo de normas como xsl o css). Además, con el fin de facilitar la gestión documental de estos nuevos recursos y controlar su heterogeneidad, se ha propuesto el diseño de herramientas que faciliten reconocer, comparar y combinar recursos web con diferente estructura: las ontologías (siendo la principal recomendación del *W3C* para su construcción el lenguaje *OWL* (*web ontology language*). Se espera que este nuevo escenario, caracterizado por la existencia de contenidos autodescritos y herramientas automáticas capaces de comprenderlos, facilite el proceso de recuperación de información y, entre otras cosas, termine con las estrategias fraudulentas de posicionamiento web.

La web semántica

Berners-Lee (2001) publicó un artículo programático en el que anunciaba el proyecto de la web semántica como una extensión de la actual, dotada de una estructura que permitiera expresar el contenido de las páginas de una forma que los ordenadores pudieran "entenderlas" y que posibilitase tanto la interacción entre ordenadores como entre éstos y los usuarios. Proponía así un nuevo modelo en el que todos sus contenidos estarían descritos y estructurados de un modo que las máquinas podrían comprenderlos.

Para que ello fuera posible, **Berners-Lee** suponía que en la web de un futuro cercano los ordenadores tendrían acceso a información semánticamente marcada y estructurada, a ontologías que expresarían conceptos, y a conjuntos de reglas de inferencia útiles para llevar a cabo razonamientos automáticos sobre las páginas web que permitiesen a los ordenadores desarrollar tareas inteligentes.

Ahora bien, de acuerdo con las previsiones iniciales este panorama descrito en el 2001 debería empezar a hacerse evidente siete años después. Tal vez porque esta transformación no ha tenido lugar, el *W3C* (que no olvidemos está dirigido por **Berners-Lee**) presenta ahora una visión mucho más prudente, orientada hacia la codificación semántica de los documentos y a la aplicación de nuevas tecnologías y procedimientos de representación del conocimiento con el fin de mejorar el acceso a los recursos de la web. Muchos de ellos se muestran a continuación.

2. Tecnologías

En sólo diez años el *W3C* ha elaborado más de ochenta especificaciones técnicas para la implantación de esta nueva infraestructura. Los principales medios con los cuales se persiguen los objetivos de la web semántica son, a grandes rasgos, los siguientes: en primer lugar, mediante una codificación de páginas en la cual las etiquetas transporten una carga semántica. Este apartado corresponde al estándar denominado xml (*Xml*, 2004). En segundo lugar, aportando descripciones (metadatos) (**Rovira**, 2006) de las páginas y sitios web con un formato que sea compatible con la estructura general de la www y con diversas categorías de páginas, e interoperable entre distintos sistemas informáticos. De esto se ocupa la norma rdf (*Rdf*, 2004). Además, mediante un sistema de ontologías que permitan especificar conceptos de diversos dominios del conocimiento mediante el uso de un lenguaje fuertemente

Término lingüístico	Sujeto	Predicado	Objeto
Término lógico	Recurso	Propiedad	Valor
Ejemplo	HpDeskjet9800	Tipo de impresión	Inyección térmica de tinta

Tabla 1: Equivalencias logicolingüísticas en una declaración rdf

basado en lógica simbólica y susceptible, por tanto, de ser eventualmente "interpretado" por un ordenador. De este aspecto se ocupa el denominado *Owl* (*Owl*, 2004). Cada una de estas tecnologías ha sido definida por varias especificaciones, y constituyen la base sobre la que el *W3C* pretende construir la web semántica.

2.1. Xml

Es sin ninguna duda el elemento de la web semántica que mayor repercusión tiene ya en biblioteconomía/documentación. Es un estándar que, junto con su norma asociada *Xml schema*, permite definir tipos de documentos y los conjuntos de etiquetas necesarias para codificarlos. La idea es que, una vez están marcados o codificados con una colección de etiquetas xml, es posible procesarlos y explotarlos de forma automática con diversos propósitos, de la misma manera que un grupo de registros de una base de datos se puede emplear de formas diversas, e incluso exportarse a diferentes sistemas de gestión de bases de datos si la estructura de registros sigue algún tipo de estándar.

"Xml es el elemento de la web semántica que mayor repercusión tiene ya en biblioteconomía/documentación"

Posibilita así a sus usuarios añadir una estructura arbitraria a sus documentos, pero sin decir nada sobre el significado de la misma, por lo que se puede considerar un meta-lenguaje para la definición de estructuras textuales.

2.2. Rdf

Es el sistema que permite utilizar metadatos para describir recursos (típicamente sitios web) en la web semántica. El objetivo de esta recomendación es habilitar la extracción del significado de la estructura de un documento, descrita en xml, con el fin de garantizar la interoperabilidad entre aplicaciones sin necesidad de intervención humana (**Senso**, 2003).

Todo el sistema rdf parte de tres entidades lógicas:

– Recursos.

– Propiedades.

– Valores.

Que se corresponden con los elementos de la lingüística:

– Sujeto.

– Predicado.

– Objeto.

Con los tres elementos anteriores podemos formar declaraciones sobre los recursos del tipo: el recurso X tiene la propiedad Y con valor P. La tabla siguiente (tabla 1) expresa las equivalencias de los componentes básicos de rdf.

Los recursos pueden ser sitios o páginas web, pero también cosas que no están en la www, como personas o cualquier objeto del mundo real o conceptual. Las propiedades son las características relevantes de los recursos (por ejemplo, con relación a las páginas web: el autor y el idioma). Por último, los valores son los datos en los que se concreta un atributo determinado de un recurso determinado. La tabla 2 expresa las ideas anteriores con dos ejemplos específicos aplicados a la descripción de dos sitios web utilizando *Dublin core*.

http://www.dublincore.org

http://www.imdb.com	dc.title	Internet movie database
http://allmovie.com	dc.title	All movie guide

Tabla 2: declaración rdf sobre dos sitios web

De acuerdo con la tabla anterior, hemos descrito dos recursos (en este caso dos bases de datos cinematográficas) mediante una de sus propiedades, concretamente el título de la página web. Para que un ordenador pueda entender este tipo de estructuras, denominadas triples, será necesario representar dicha información mediante rdf/xml y *Dublin core*. En la figura 1 mostramos esta representación para uno de los triples que aparecen en la tabla 2.

Ignoramos cómo evolucionará rdf pero, afortunadamente, la amplitud de miras de esta recomendación no es un obstáculo para su aplicación al mundo de la documentación, sino todo lo contrario: una de sus más importantes y significativas utilidades consiste en la descripción de recursos digitales (**Rovira**, 2007) utili-

```
<?xml version="1.0" ?>

<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:dc="http://purl.org/dc/element/1.1">

 <rdf:Description rdf:about="http://allmovie.com">

 <dc:title>All movie guide</dc:title>

 </rdf:Description>

</rdf:RDF>
```

Figura 1: Serialización de uno de los triples de la tabla 2

zando *Dublin core*, norma que, como es sabido, consiste precisamente en aplicar la filosofía documental a la descripción de recursos.

2.3. Owl

Con el objeto de que las máquinas puedan realizar tareas de razonamiento útil sobre los recursos de la web semántica, es necesario definir un lenguaje o herramienta de descripción que vaya más allá de las semánticas básicas de rdf y permita comparar y combinar documentos (recursos) con distinta estructura (es decir, que sea capaz de reconocer, por ejemplo, el elemento <*proveedor*> y <*provider*> de dos estándares para la gestión de transacciones comerciales como iguales, permitiendo la interoperabilidad entre ambos). A estos lenguajes o herramientas se les denomina ontologías, y básicamente incluyen las definiciones de los conceptos, denominadas "clases", de un dominio y las relaciones entre ellos.

Owl es el lenguaje estándar de la web semántica para expresar y codificar ontologías. Por tanto, puede ser utilizado para representar explícitamente el significado de términos en vocabularios y las relaciones (semánticas) entre ellos.

http://www.w3.org/TR/2004/REC-owl-features-20040210/

Consigue formalizar las relaciones entre las clases aún más que rdf, indicando aspectos básicos para el razonamiento como la existencia de conceptos o clases disjuntas en un dominio. Por ejemplo, "los periféricos de salida no son periféricos de almacenamiento", esto es, la clase de los periféricos de salida es disjunta a la clase de los de almacenamiento. También es posible expresar la cardinalidad, es decir, el número de elementos que pueden componer un concepto o clase, por ejemplo, "un libro puede tener uno o varios autores" (la cardinalidad de los autores de un libro es uno o más de uno), o bien "un libro solamente puede tener un isbn" (la cardinalidad del isbn de los libros es exactamente uno). Puede expresar igualdad o equivalencia entre clases, características y restricciones de las mismas, etc.

Owl utiliza rdf para representar y codificar las ontologías. Esta recomendación sigue la tendencia tan característica del *W3C* de proceder mediante "extensiones". Por tanto, owl es una extensión de rdf que añade elementos como los mencionados anteriormente para describir características y clases.

A modo de ilustración, en la figura 2 podemos ver un gráfico que representa un ejemplo de clases y subclases de una ontología de periféricos de ordenador:

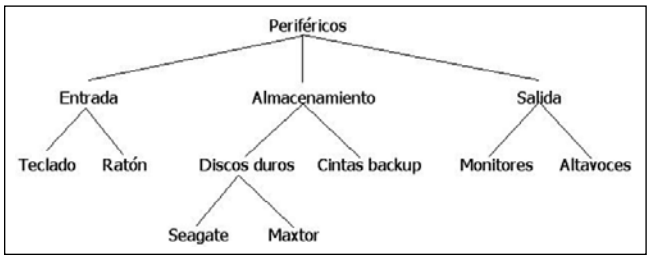


Figura 2: Clases y subclases de una ontología sobre periféricos de ordenador (Codina, 2006)

A continuación, en la figura 3, vemos parte de la ontología anterior representada mediante owl.

```
<?xml version="1.0" ?>

<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
 ...
<owl:Class rdf:ID="Perifericos">
 <rdf:comment>
Los periféricos de ordenador están conectados a la CPU pero no forman parte de ella.
 </rdf:comment>
</owl:Class>

<owl:Class rdf:ID="Entrada">
 <rdf:comment>
Los periféricos de entrada son una subclase de periféricos de ordenador.
 </rdf:comment>

 <rdfs:subClassOf rdf:resource="#Perifericos" />
</owl:Class>

<owl:Class rdf:ID="Teclados">
 <rdf:comment>
            Los teclados son una subclase de los periféricos de entrada.
 </rdf:comment>

 <rdf:subClassOf rdf:resource="#Entrada" />

 <rdf:subClassOf rdf:resource="#Perifericos" />
</owl:Class>
...
 </rdf:RDF>
```

Figura 3: Representación en owl de la ontología de la figura 2

La idea esencial es que en algún momento del futuro (pero no parece probable que sea a corto plazo), la web no solamente estará poblada por un determinado número de ontologías que permitirán a los ordenadores realizar inferencias sobre la información publicada, sino que los agentes de usuario (esto es, los navegadores del futuro) serán capaces de realizar razonamientos fiables sobre tales ontologías. No obstante, el coste

requerido en tiempo y dinero para la generación de la infraestructura propia de la web semántica, la ausencia de alicientes que animen a los usuarios a adoptar las recomendaciones del *W3C* para el desarrollo de sus contenidos, y el desinterés de parte de los agentes técnico-comerciales más importantes de la web actual (como por ejemplo los proveedores de servicios de búsqueda y los fabricantes de navegadores), hacen augurar que incluso la fecha del 2010 que algunas veces se ha apuntado para este escenario resulte optimista.

Afortunadamente, entendemos que no es necesario esperar que se haga realidad (si es que alguna vez sucede) lo que podríamos denominar el "lado visionario de la web semántica" representado por el artículo de **Berners-Lee** del año 2001. Las ontologías son una tecnología que pueden aportar soluciones ahora a problemas reales actuales.

3. Ontologías: la nueva visión

En los últimos años, el proyecto de la web semántica ha servido para asentar lo que podríamos denominar el uso actual del término ontología (no olvidemos que ha gozado de varias vidas desde sus orígenes en la filosofía clásica). En este nuevo contexto, una de las definiciones más citadas es la de **Studer** (1998) -que completa la original de **Gruber** (1993)-, para quien una ontología es "una especificación explícita y formal de una conceptualización compartida". Una "conceptualización" es un modelo abstracto de algún fenómeno del mundo construido mediante la identificación de los conceptos relevantes a ese fenómeno (normalmente un dominio del conocimiento). "Explícito" significa que los conceptos utilizados en la ontología, y las restricciones para su uso, están claramente definidos. "Formal" se refiere al hecho de que debe ser comprensible para las máquinas, es decir, estar expresada mediante una sintaxis (como owl) que permita a un ordenador operar sobre ella. Por último, "compartida" refleja la noción de que contendrá conocimiento consensuado en algún grado (en el caso de un dominio del conocimiento, se supone que estará consensuado por los expertos en él) (**Gómez-Pérez**, 2005).

En este contexto, para que la especificación de un dominio se considere una ontología, debe presentar, al menos, dos tipos de componentes: elementos y relaciones entre los mismos. Los primeros son los siguientes (**Lacy**, 2005, pp. 32-40):

– Clases: las entidades del "mundo real" se pueden categorizar en grupos o conjuntos de objetos con similares características, forman las clases de la ontología. Las entidades pueden ser cosas físicas (p. e. automóviles) o conceptuales (p. e. teorías científicas). Algunos serían "País", "Libro" y "Automóvil". Constituyen el núcleo de una ontología y describen los conceptos de un dominio concreto. Como hemos señalado, un ejemplo de una clase podría ser "Automóvil", que idealmente representaría a todos los automóviles del mundo. De este modo, cada coche que vemos por la calle es una instancia o ejemplar de la clase "Automóvil". Una clase puede tener además subclases, que representan conceptos más específicos que el de su superclase. De este modo, podríamos dividirla por ejemplo en las subclases "Turismo", "Todoterreno" y "Deportivo".

– Propiedades: las entidades que pertenecen a una clase poseen atributos determinados, p. e. tienen un nombre, un color o un peso. Por tanto, las propiedades consisten en pares de atributo/valor y sirven para describir de forma conveniente las características relevantes de las entidades que forman las clases. Algunos ejemplos son "Población", "Isbn" y "Precio".

– Individuos, instancias o ejemplares: consisten en representaciones de objetos o elementos particulares de una clase. Se denominan indistintamente individuos o instancias de la clase. Hay que señalar que es difícil (y de hecho es discrecional) distinguir entre individuos y clases. Ejemplos de instancias son "España", "Documento 141203448-5" y "Ford Mustang".

Por su parte, las relaciones típicas de una ontología son las siguientes:

– Clase–Individuo: asocian individuos o instancias a una clase. Por ejemplo, "España" es una instancia de la clase "País", "Ford Mustang" lo es de "Automóvil" y "Documento 141203448-5" de "Libro". Se expresan mediante relaciones "es un" ("*is a*"), por ejemplo, "Ford Mustang es un automóvil".

– Individuo–Propiedad: como hemos señalado anteriormente, las instancias de una clase tienen valores asociados a propiedades. Estas asociaciones se expresan mediante relaciones "tiene el valor" ("*has value for*"). Por ejemplo, España tiene el valor "505 mil km2" para la propiedad "Extensión".

– Clase–Propiedad: la clase como un conjunto tiene propiedades. Cuando se aplican a una clase, estas propiedades se denominan restricciones porque sirven tanto para definirla como para delimitar la pertenencia de los individuos a ella. Por ejemplo, "Automóvil" posee la propiedad "tener un motor", que excluye de la misma a los vehículos de tracción animal.

– Clase–Subclase: las clases pueden tener subclases. Por ejemplo, "Todoterreno" es una subclase de "Automóvil". Esta asociación se expresa también con relaciones "es un".

Además de las anteriores, en una ontología también se dan otras clases de relaciones atendiendo a otros enfoques. En concreto, se suelen contemplar relaciones entre conceptos (clases) de sinonimia, antonimia, hipo-

nimia[1] y meronimia[2]. Algunas son similares a las que se contemplan en los tesauros. Además, cabe recordar que las relaciones clase-subclase y clase-individuo son la base de taxonomías y tesauros, de aquí la tendencia ya señalada a confundir las tres cosas.

"Las ontologías pueden considerarse lenguajes documentales con distintos niveles de estructura, pero a diferencia del tesauro tradicional están elaboradas con una sintaxis comprensible para los ordenadores"

Todas estas similitudes no deberían hacernos caer en el error de concebir una ontología como un tesauro (o como una taxonomía). Ciertamente, y al igual que un tesauro, las ontologías pueden considerarse lenguajes documentales con distintos niveles de estructura, pero a diferencia del tesauro tradicional, en primer lugar, están elaboradas con una sintaxis comprensible para los ordenadores. Además, como hemos visto, las ontologías contemplan un conjunto más amplio de relaciones que las de clase y subclase (como en una taxonomía) o las de sinonimia y meronimia (como en un tesauro) ya que en principio estas relaciones no están cerradas, sino que en parte dependen de las relaciones reales que se den entre las clases y los individuos del dominio modelado por la ontología. Por tanto, una ontología permite mayor riqueza en la definición de sus conceptos y sus relaciones que un tesauro.

Sin embargo, la diferencia más importante es el hecho de que están formalizadas, es decir expresadas mediante una rigurosa lógica formal y, por tanto, no solamente pueden ser procesadas por aplicaciones informáticas sino que, en principio (aunque con severas limitaciones) soportan procesos de inferencia automáticos.

3.1. Generación de ontologías

Acudiendo ahora a términos mucho más prácticos, y siguiendo la metodología especialmente relevante que propone **Noy** (2001), a grandes rasgos, el desarrollo de una ontología implica, al menos las siguientes fases:

– Definir las clases (conceptos).

– Ordenar las clases en una taxonomía.

– Definir las propiedades de las clases y los valores asociados a esas propiedades.

– Completar los valores de las propiedades para cada una de las instancias reales.

Debe tenerse en cuenta que no existe un modo correcto de modelar un dominio: siempre encontraremos distintas alternativas para hacerlo que nos proporcionarán diferentes resultados. Obsérvese que esta afirmación conlleva la concepción de las ontologías como instrumentos adaptados a la resolución de tareas, y por ende, la concepción de las mismas como conceptualizaciones "no" universales de los dominios que representan, lo cual hoy por hoy, choca frontalmente con la concepción universalista de las ontologías del *W3C*. Por tanto, el diseño de una ontología estará condicionado por su uso y nivel de detalle.

En cuanto a la complejidad asociada a su elaboración, el primer problema es determinar qué términos debemos enunciar y qué propiedades vamos a enumerar de éstos. Para solucionarlo es muy importante obtener a priori una lista de los términos que consideremos relevantes al dominio, sin preocuparnos de si existe solapamiento entre sus significados, o de las relaciones entre ellos. Las dos siguientes etapas serán desarrollar la jerarquía de clases (la taxonomía) y definir las propiedades de los conceptos.

Existen distintas aproximaciones para extraer la taxonomía de clases. Podemos recurrir a una aproximación arriba-abajo (*top-down*), que comienza con la definición de los conceptos más generales en el dominio para a continuación extraer conceptos más específicos. O por el contrario utilizar una metodología de abajo-arriba (*bottom-up*), mediante la identificación de las clases más específicas, y a continuación la agrupación de éstas en otras clases más generales. Aunque también se puede adoptar una aproximación mixta que combine los dos enfoques anteriores. En este caso, en primer lugar se identifican los conceptos más relevantes para el dominio de la ontología, y a continuación se generalizan o especializan según sea conveniente. No puede afirmarse que uno u otro método sea más apropiado para la extracción de la taxonomía, así que la selección del mismo dependerá de nuestra percepción del dominio.

3.2. Técnicas para la generación semiautomática de ontologías

Se espera que la aplicación de las especificaciones del *W3C*, junto con el desarrollo y generalización de las ontologías suponga el final de los problemas derivados de la ausencia de un modelo de datos bien definido en la web. No obstante, para que esto sea posible hay que solventar un nuevo problema, a saber: la estructuración y descripción de los recursos web mediante xml y rdf, así como que la elaboración manual de ontologías supone un coste tan elevado, en tiempo y dinero, que ya son muchas las voces que cuestionan que la transformación de la web actual en la web semántica pueda llegar a ser una realidad algún día.

"Muchos dudan de que la transformación de la web actual en la web semántica pueda llegar a ser una realidad algún día"

Con el fin de intentar paliar este problema ha aparecido una nueva disciplina, la ingeniería de ontologías, dedicada al estudio y diseño de aplicaciones que ayuden a su elaboración, mantenimiento y uso. Su principal objetivo es, por tanto, la creación de entornos que, mediante la automatización de ciertas tareas y el diseño del software para su gestión, agilicen el proceso. Ejemplos:

– *KAON*.
http://kaon.semanticweb.org/

– *Hozo*.
http://www.hozo.jp/

– *WebODE*.
http://webode.dia.fi.upm.es/WebODEWeb/index.html

– *Protégé*.
http://protege.stanford.edu/

Una comparación de estos sistemas puede encontrarse en **Mizoguchi**, 2004.

Mención especial merece la disciplina conocida como "Aprendizaje de ontologías" u "*Ontology learning*" (**Maedche**, 2004), una parte de la ingeniería de ontologías que investiga el desarrollo de métodos para la creación de una ontología de forma semiautomática. Concretamente, se centra en la generación de herramientas que permitan importar, extraer, podar, refinar y evaluar la taxonomía de una ontología bajo la supervisión de un experto humano, el "ingeniero ontológico", denominación acuñada en el ámbito germano, y cuyo perfil se corresponde en gran medida con el de un documentalista.

A continuación, para ilustrar el funcionamiento de estos sistemas se describe brevemente la arquitectura de una de las primeras propuestas formuladas en este ámbito (figura 4), la de **Maedche** (2001), que ha determinado las líneas básicas a seguir en este campo. La arquitectura propuesta consta de cuatro elementos:

– Interfaz gráfica: permite al ingeniero ontológico intervenir manualmente en todo el proceso de creación.

– Componente de gestión: con ella seleccionamos los datos a partir de los cuales construir la ontología (documentos html y xml, dtds, bases de datos, otras ontologías, etc.).

– Centro de procesamiento de recursos: facilita al ingeniero ontológico diferentes herramientas para procesar los documentos de entrada y extraer la terminología necesaria (conceptos).

– Por último, el sistema (véase *KAON*) dispone de una biblioteca de algoritmos cuyo funcionamiento se basa normalmente en reglas de asociación, técnicas de análisis formal de conceptos, o técnicas de agrupamiento (jerárquicas o no). Mediante la aplicación de uno o varios de estos algoritmos sobre los documentos ya procesados podrán extraerse las clases de la taxonomía y sus relaciones.

En teoría, mediante el uso de un sistema como el descrito, podría construirse una ontología siguiendo las siguientes etapas:

– Si es posible, importamos y reutilizamos las ontologías existentes en el dominio de nuestro interés, y un experto las fusiona (manual o automáticamente) en una única a partir de la cual aplicar el resto de fases.

– Extracción de ontologías: la herramienta propone diferentes entradas léxicas (términos) para la ontología, que se obtienen en función del procesamiento de los textos de los recursos del dominio seleccionados (documentos html, etc.). Independientemente de la recomendación hecha por el sistema, el ingeniero ontológico puede incluir o eliminar entradas léxicas si así lo desea. Obtenido el léxico, el siguiente paso es su clasificación taxonómica mediante técnicas automáticas de clasificación. El resultado final de esta fase es la propuesta de una taxonomía del dominio al ingeniero ontológico que éste puede modificar o rehacer como crea conveniente.

– Poda de ontologías: la arquitectura viene dotada de herramientas que permiten al ingeniero ontológico ajustar la ontología a su propósito original.

– Refinamiento: el sistema pone a disposición del profesional herramientas que permiten completar y afinar el resultado final.

– Evaluación de la ontología resultante: a través del seguimiento y observación de su uso.

– Actualización: para incluir nuevos dominios o actualizar los ya existentes.

Este modelo asume que cualquier ontología puede ser descrita por un conjunto de conceptos, relaciones y entradas léxicas. Consecuentemente, su construcción se puede agilizar utilizando tecnologías que analicen distintos tipos de recursos web (documentos html, xml, dtds, bases de datos, etc.) y extraigan los términos más significativos para nuestro dominio de interés así como sus relaciones. Todo ello bajo la supervisión de un experto humano en generación de ontologías.

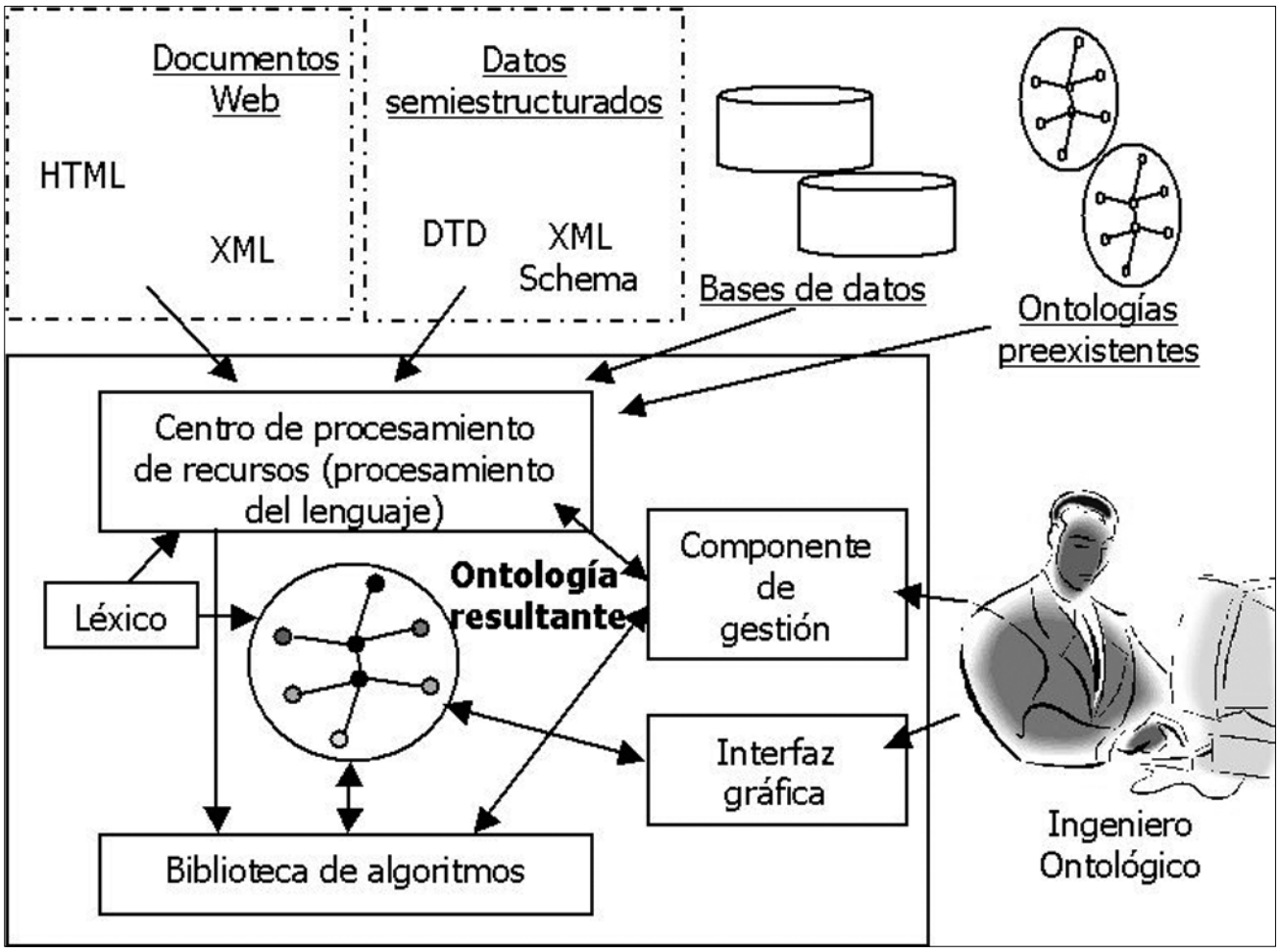


Figura 4: Arquitectura del sistema para el aprendizaje de ontologías propuesto por Maedche (2001)

4. Conclusiones

A la luz de todo lo expuesto es inevitable plantearse ciertas cuestiones que determinarán en gran parte el futuro. La primera de ellas quizás sea: cuando nos pronunciamos acerca de la web semántica ¿estamos hablando de una realidad palpable? La respuesta a esta pregunta es muy compleja. Si por realidad entendemos la existencia y funcionamiento de este entorno en la actualidad, puede afirmarse que no lo es.

En cambio, sí son realidades las iniciativas de investigación y desarrollo (públicas y privadas) puestas en marcha a raíz de la formulación de esta nueva Web y las recomendaciones del *W3C* como el lenguaje xml y rdf que están influyendo de forma activa y real en buena parte de la Web actual.

¿Significa esto que la web semántica será, entonces, una realidad en el futuro? Cada vez más analistas creen que es poco probable que se haga realidad el lado más visionario del proyecto, ni a corto ni a medio plazo.

De lo que no cabe duda es que aportará muchas cosas por el camino y provocará cambios duraderos y decisivos que ayudarán a tener una Web mucho mejor en el futuro. Un ejemplo fácil lo tenemos en la potente idea de la separación entre presentación y contenido de los documentos web.

¿Qué metodología se impondrá para la generación de esta futura Web: el procesado manual o el semiautomático? Por un lado, las descripciones de los contenidos y las ontologías elaboradas por expertos humanos son de gran calidad, aunque su coste en tiempo y dinero es inabarcable (además cabe la posibilidad de fraude); por otro, la utilización de herramientas automáticas para agilizar el desarrollo de las descripciones y las ontologías supone una disminución considerable de los costes, a cambio de descripciones y ontologías más someras (a menudo meras taxonomías o clasificaciones) que, además, pueden ser erróneas, y que difícilmente satisfacen las exigencias mínimas de las especificaciones del *W3C*.

Finalmente, quizás lo más probable (y apropiado) consista en una aproximación mixta, es decir, la coexistencia de descripciones y ontologías manuales junto con las automáticas. Las primeras, utilizadas cuando el dominio o tarea requiera descripciones y ontologías de gran calidad, y se disponga de los recursos necesarios. Las segundas, cuando el dominio o la tarea en cuestión suponga una tarea inabarcable para un experto

humano, bien por su magnitud, bien por su naturaleza cambiante.

En todo caso es de esperar que los logros aportados por este nuevo entorno web sean adecuadamente incorporados a cualesquiera otros escenarios dedicados a la gestión de la información documental. En particular, el profesional y el estudioso de la biblioteconomía-documentación no debería quedar al margen de estos avances.

"Es de esperar que los logros de este nuevo entorno web sean adecuadamente incorporados a la gestión de la información documental"

De hecho, la formación y experiencia de esta clase de profesionales hacen de ellos firmes candidatos a jugar un papel preferente en el desarrollo de la web semántica. Especialmente útil sería su participación activa en todas aquellas tareas conducentes a la descripción de los recursos de esta nueva web, poniendo especial énfasis en la importancia de las ontologías.

Probablemente el campo emergente de la ingeniería de ontologías favorecerá las aproximaciones semiautomáticas, asignando un papel preeminente al experto humano, el ingeniero ontológico, en el desarrollo de estas herramientas. Por ahora este nuevo perfil, muy similar al de un documentalista, parece vinculado exclusivamente a la informática; pero nada impide que, dada la similitud indicada, los profesionales e investigadores de nuestro campo tengamos también un papel más o menos protagonista. Existen, de hecho, otros campos profesionales y científicos que podríamos denominar "compartidos". En la arquitectura de la información, por ejemplo, podemos encontrar tanto a profesionales de la biblioteconomía-documentación como de la informática (y casos que comparten ambos perfiles, claro).

Corresponde a los documentalistas y profesionales de la información hacer visible su idoneidad para el desempeño de estas nuevas labores. Como expertos en la elaboración de lenguajes documentales y herramientas para el control terminológico deberían ser un agente más en la creación de esta nueva Web, evaluando tanto las recomendaciones como las herramientas para la descripción y recuperación de los nuevos recursos web, y asesorando a aquellos involucrados en su diseño. Sin duda alguna, los documentalistas y profesionales de la información están hoy en situación de adquirir las habilidades técnicas que les permitan desempeñar estas labores eficazmente. Si así lo hacen, probablemente asistiremos al nacimiento de una nueva dimensión que puede revalorizar considerablemente el perfil de este profesional de la información.

Notas de la Redacción

1. En semántica lingüística, se denomina hipónimo a aquella palabra que posee todos los rasgos semánticos, o *semas*, de otra más general, su hiperónimo, pero que añade en su definición otros rasgos semánticos que la diferencian de la segunda.
Ejemplo: *lunes*, *martes*, *miércoles*, etc. son hipónimos de *día*.
Fuente: *Wikipedia* en español.

2. La meronimia es una relación semántica no-simétrica entre los significados de dos palabras dentro del mismo campo semántico. Se denomina merónimo a la palabra cuyo significado constituye una parte del significado total de otra palabra, denominada ésta holónimo. Por ejemplo, *dedo* es merónimo de *mano* y *mano* es merónimo de *brazo*; a su vez, *brazo* es holónimo de *mano* y *mano* es holónimo de *dedo*.
Ejemplos: *azul* es merónimo de *color*; *doctor* es merónimo de *oficio*.
Fuente: *Wikipedia* en español.

Agradecimientos

Este trabajo ha sido financiado por el *Ministerio de Educación y Ciencia*, como parte del proyecto *HUM2004-03162/FILO*.

Bibliografía

Berners-Lee, T.; **Hendler, J.**; **Lassila, O.** "The semantic web". En: *Scientific American*, 2001, May, v. 284, n. 5, pp. 34-43.

Codina, L.; **Rovira, C.** "La web semántica". En: **Tramullas, J.** (ed.). *Tendencias en documentación digital*. Gijón: Ediciones Trea, 2006, pp. 9-54.

Extensible markup language (xml) 1.1 (W3C Recommendation 04 Feb 2004, edited in place 15 Apr 2004).
http://www.w3.org/TR/2004/REC-xml11-20040204/

Gómez-Pérez, A.; **Manzano-Macho, D.** "An overview of methods and tools for ontology learning from text". En: *The knowledge engineering review*, 2005, v. 19, n. 3, pp. 187-212.

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

Maedche, A.; **Staab, S.** "Ontology learning for the semantic web". En: *IEEE intelligent systems*, 2001, v. 16, n. 2, pp. 72-79.

Maedche, A.; **Staab, S.** "Ontology learning". En: **Staab, S.**; **Studer, R.** (eds.). *Handbook on ontologies*. Berlin: Springer, 2004, pp. 173-189.

Mizoguchi, R. "Ontology engineering environments". En: **Staab, S.**; **Studer, R.** (eds.). *Handbook on ontologies*. Berlin: Springer, 2004, pp. 275-296.

Noy, N. F.; **McGuinness, D. L.** "Ontology development 101: a guide to creating your first ontology". En: *Stanford Knowledge Systems Laboratory Technical report KSL-01-05*.

OWL Web Ontology Language: Overview (W3C Recommendation 10 Feb 2004).
http://www.w3.org/TR/owl-features/

RDF Vocabulary description language 1.0: RDF Schema (W3C Recommendation 10 Feb 2004).
http://www.w3.org/TR/rdf-schema/

Rovira, C.; **Marcos, M. C.** "Metadatos en revistas–e de documentación de libre acceso". En: *El profesional de la información*, 2006, marzo-abril, v. 15, n. 2, pp. 136-144.

Rovira, C.; **Marcos, M. C.**; **Codina, L.** "Repositorios de publicaciones digitales de libre acceso en Europa: análisis y valoración de la accesibilidad, posicionamiento web y calidad del código". En: *El profesional de la información*, 2007, enero-febrero 2007, v. 16, n. 1, pp. 24-38.

Senso, J. A. "Herramientas para trabajar con rdf". En: *El profesional de la información*, 2003, marzo-abril, v. 12, n. 2, pp. 132-139.

Studer, S.; **Benjamins, R.**; **Fensel, D.** "Knowledge engineering: principles and methods". En: *Data and knowledge engineering*, 1998, n. 25, pp. 161-197.

***Rafael Pedraza-Jiménez**, Área de Biblioteconomía y Documentación, Universidad Pompeu Fabra, Barcelona.*
rafael.pedraza@upf.edu

***Lluís Codina**, Facultad de Comunicación Audiovisual, Universidad Pompeu Fabra, Barcelona.*
lluis.codina@upf.edu

***Cristòfol Rovira**, Área de Biblioteconomía y Documentación, Universidad Pompeu Fabra, Barcelona.*
cristofol.rovira@upf.edu

On the nature and typology of documentary classifications and their use in a networked environment

Por **Aida Slavic**

Resumen: *Los estándares para publicar e intercambiar vocabularios enfocados en la Red, así como las propuestas para servicios y registros de terminología, pueden mejorar el intercambio y uso de todos los sistemas para organizar conocimientos en red. Esto significa que las clasificaciones documentales también pueden resultar más útiles fuera de su aplicación original. En este trabajo se resumen unas características típicas de clasificaciones documentales y se explican aspectos de terminología, función y realización. El concepto original de cada esquema de clasificación determina las funciones para las que el vocabulario está diseñado. Estas funciones influyen sobre la estructura, semántica y sintáctica, así como la cobertura de esquemas y formato en el cual los datos de clasificación se publican y se hacen accesibles. El autor sugiere que hay que prestar atención a las diferencias entre clasificaciones documentales porque pueden determinar su encaje para un fin específico y pueden imponer requisitos distintos a su uso online. Cuando hablamos creamos muchas clasificaciones para organizar conocimientos y puede ser importante promocionar una mayor pericia del dominio bibliográfico respecto a la construcción y uso de sistemas de clasificación.*

Palabras clave: *Esquemas de clasificación, Clasificación documental, Clasificación bibliotecaria, Tipología*

Título: Características y tipología de clasificaciones documentales y su uso en un ámbito de red

Abstract: *Networked orientated standards for vocabulary publishing and exchange and proposals for terminological services and terminology registries will improve sharing and use of all knowledge organization systems in the networked information environment. This means that documentary classifications may also become more applicable for use outside their original domain of application. The paper summarises some characteristics common to documentary classifications and explains some terminological, functional and implementation aspects. The original purpose behind each classification scheme determines the functions that the vocabulary is designed to facilitate. These functions influence the structure, semantics and syntax, scheme coverage and format in which classification data are published and made available. The author suggests that attention should be paid to the differences between documentary classifications as these may determine their suitability for a certain purpose and may impose different requirements with respect to their use online. As we speak, many classifications are being created for knowledge organization and it may be important to promote expertise from the bibliographic domain with respect to building and using classification systems.*

Dr. ***Aida Slavic*** *is one of the associate editors of the Universal Decimal Classification (UDC Consortium) and visiting lecturer at the Department of Information Sciences at the University of Zagreb, Croatia. Dr. Slavic has undergraduate and postgraduate degrees in library and information science from University of Zagreb. In 2005 she was awarded a PhD in library and information studies from University College London. She participated in several international and UK projects in the area of resource discovery metadata and controlled vocabularies. Her research interests are in the use of classification in resource discovery and modelling and formatting of controlled vocabularies for their use in a networked environment.*

Keywords: *Classification schemes, Documentary classifications, Library classifications, Typology.*

Slavic, Aida. "On the nature and typology of documentary classifications and their use in a networked environment". En: *El profesional de la información,* 2007, noviembre-diciembre, v. 16, n. 6, pp. 580-589.

DOI: 10.3145/epi.2007.nov.05

1. Introduction: information organization and subject indexing

With the development of information and communication technology we are more exposed to problems of information overload. At the same time, we have gained efficient ways of finding information and are in a better position of combining and exploiting a range of information organization tools and methods. In a networked environment we deal with information organization problems in various contexts and scenarios and one single approach or method of information organization does not fit all purposes. For instance, when looking for information about some topic in large digital text collections, we may rely, to a certain extent, on powerful text retrieval and relevance ranking

Artículo recibido el 06-08-07
Aceptación definitiva: 15-10-07

techniques. It is, however, much more difficult to find relevant items in integrated digital collections containing heterogeneous resources (data sets, images and sound) or finding information that is scattered in digital, printed or realia collections. Hence, we still need some "traditional methods" in information organization such as metadata and human content indexing and we ought to combine these with new and emerging approaches, especially in the context of an integrated access to information.

Apart from new textual digital resources that are constantly emerging on the internet and can be accessed, with more or less precision, using general searching services (search engines), we also have to deal with legacy data. For instance, ready-indexed library collections worldwide are now becoming more exposed and open for cross searching. Hundreds of millions of these information resources are already organized using some traditional knowledge organization systems based on literary warrant and accepted scientific and educational consensus such as documentary classifications, thesauri or subject-heading systems. At the same time we have large collections of digital-born material indexed, i.e. tagged, by users in the context of web 2.0, for instance, social bookmarking services such *Digg*, *Del.icio.*us, *Flickr*, *Simpy*, *Connotea* etc., that organize information with strong emphasis on the way users view and access these resources (user warrant).

It is only natural, that libraries, for instance, will be looking now into ways of combining and making use of various information organization and information discovery methods, such as social tagging and internet search engines, to extend user services provided through library catalogues. In doing so they may review the way in which they currently exploit their own knowledge organization tools in providing subject access to information-and may need to know which knowledge organization function is valuable to users and cannot be replaced by other methods.

Over time, library and information professionals have developed various subject indexing systems. A feature common to these systems is the provision of methods and techniques for controlling the ambiguity and vagueness of natural language and thus allowing for document content to be described in an unambiguous and more standardized way. Sometimes information professionals refer to these systems as controlled vocabularies in order to emphasise that the vocabulary used to describe documents consists of a selection of terms in which the problems of homonymy and synonymy are resolved. Librarians prefer to use the expression subject indexing languages which emphasises the fact that a document indexing system can operate with both a controlled vocabulary and syntax rules prescribing how the terms may be combined in complex subject statements that resemble a language. Subject indexing systems perform the following functions: they unambiguously define the meaning of concepts within a knowledge system; and they relate concepts according to the level of their semantic closeness/similarity. Subject indexing languages are usually categorised in some of the following ways:

a. According to the type of terms they use to represent concepts:

– Alphabetical subject indexing languages which use natural language as indexing terms, such as subject-heading systems and thesauri or any other indexing system using words.

– Classifications using symbols as indexing term such as documentary classifications.

b. According to the type of relationships they can express between concepts:

– Expressing only semantic i.e. hierarchical and associative relationships, such as thesauri, classifications.

– Expressing only syntactic relationships such as subject-heading systems and some less widely known systems such as *PREserved Context Indexing System-Precis* (**Foskett**, 1997) or *Postulate-based Permuted Subject Indexing-Popsi* (**Bhattacharyya**, 1979).

– Expressing both semantic and syntactic relationships (faceted or analytico-synthetic documentary classifications).

Indexing languages that provide syntax rules for combining concepts in the process of indexing are also called pre-coordinated indexing languages as opposed to post-coordinated indexing languages in which simple index terms are assigned to documents and are combined in the process of searching only (**Svenonius**, 1995).

In representing knowledge for the purpose of systematic browsing, alphabetical indexing languages will order topics or subjects alphabetically thus subjects represented with words beginning with the same letter will be collocated together irrespective of their semantic relationships. Classification systems will represent knowledge areas systematically and topics that have similar or semantically related meaning will be collocated together irrespective of what natural language terms are used to describe them. While alphabetical indexing languages have the advantage of using natural language, classifications have the advantage of supporting systematic organization. This is the reason why alphabetical and classificatory indexing languages are often combined and used in a complementary way.

Most specifically this is the case with thesauri and classifications, hence in the past we had indexing systems called thesaurofacets (**Aitchison**, 1970) or classaurus (**Bhattacharyya**, 1982).

Both thesauri and classifications express hierarchical (broader and narrower concepts) and associative relationships (related concepts) between subjects and thus they can be utilised for search expansion and improving of both recall and precision in information retrieval. In a thesaurus, for instance, the hierarchy is established between a concept and its next broader category. The same concept, however, may have different semantic relationships in different areas of knowledge. For example, "Soil" in a thesaurus of agriculture will have a different broader category than "Soil" in a thesaurus of geology. This is why thesauri are usually developed for a specific subject area. Classifications, on the other hand, have mechanisms of visually representing concepts and their relationships in an entire field or, indeed, entire universe of knowledge and are well suited for both hierarchical browsing of a single subject area and for navigating between different fields of knowledge.

1.1. The role of classification in networked environment

Classificatory vocabularies offer a plethora of solutions on how we may categorise and aggregate information resources. They may support the use and development of other indexing methods and increase their efficiency.

"Categories can alleviate indexing and render this procedure more reliable in that they control the selection of essential concepts. Furthermore, they can help to guarantee just that degree of representational predictability and fidelity which an index language displays and which one expects of it as an information system user. Without such a categorial guide an index language may well be (or soon become!) treacherous with respect to these most important system features" (**Fugmann**, 1990, p. 67).

Digital environments have widened the field of application of classifications significantly, providing a large space for the testing, implementation and comparison of different systems. Without the limitations imposed by physical documents it is possible, and sometimes desirable, to organize the same collection of resources according to different knowledge organization systems. We may index the same collection by two different classification systems and the user can choose the "view" into the collection that suits his needs. We may also map several indexing languages to a single classification that acts as a pivot or switching language in integrated services -and enable browsing and a seamless view of the collection no matter which indexing systems are used locally. Classification schemes are mapped, linked and combined with other classifications and alphabetical indexing languages. In addition, instead of the linear browsing of subjects we are now looking into improving facet-based views and coordinated navigation between different facet hierarchies. For instance, the same collection may be systematically browsed by place, time, topic, audience or by form of presentation. Users may combine facets in searching or switching from one to another when browsing the collection.

Existing and emerging standards for networked vocabulary exchange aim to provide machine readable formats for expressing classifications irrespective of their type and application purposes. Thus, developers of *ISO/IEC 13250 Topic Maps* (2000), *BS 8723 Structured vocabularies for information retrieval* (2005) and *Simple Knowledge Organization System* (*SKOS*) (2006) are all concerned with the typology of vocabularies, their function in information retrieval, data structure and data naming. Various parties and stakeholders are interested in common formats that offer a satisfactory compromise for all vocabularies including classification schemes. "Terminology services" and "terminology registries" are now more often proposed as a solution that may exploit vocabulary standards to provide an integrated and fully managed subject access control and mapping service between various vocabularies for their use and sharing in a networked environment (**Vizine-Goetz**, 2004; **Tudhope**; **Koch**; **Heery**, 2006).

Documentary classifications are a group of more complex classification systems that differ one from another on a structural, syntactic and functional level. For instance, modern analytico-synthetic and faceted classifications have greater potential in knowledge organization which requires a greater level of machine readability with respect to the management of facet hierarchies, their browsing and post-coordinate searching (**Slavic**; **Cordeiro**, 2004, 2004a). If standards are created only on those characteristics that are shared by all classifications -it is inevitable that more modern and more sophisticated documentary classifications may not be able to support all of the information retrieval functions for which they are designed. This is why it may be important to understand the structural differences between documentary classifications, how they are built and for what purpose, how they are used for indexing and how they are used in information retrieval. This paper will summarise some of the main characteristics of this particular group of classification schemes in order to highlight areas that could benefit from further exploration.

2. Classifications for knowledge mediation

In the field of information and knowledge organization we use the expression "classification" to denote logically organized, hierarchically and semantically structured schemes of concepts that are created for the purpose of document content indexing and knowledge mediation (cf. *ISO 5963*, 1985). Within classification schemes concepts are organized into classes and subclasses and each class may be, if needed, represented with a notational symbol (numeric, alphabetic or alphanumeric). Using such a notational scheme one can unambiguously denote a class/concept without using natural language. But most importantly, by browsing a well structured scheme one can discern the semantic relationships between concepts; i.e. find out what are their broader, narrower, collateral and related classes/concepts (**Bhattacharyya**, 1979).

In order to explain the relationship between documentary and other knowledge classifications, **Dahlberg** (1992) recommended Shamurin's classification of knowledge classifications. According to this framework all knowledge classification are created for one of the following purposes: knowledge representation (philosophical classification systems, education-oriented classification systems); knowledge utilization (encyclopaedic classification systems, word classification systems and linguistic thesauri); knowledge mediation (bibliographic classification systems, documentation classification systems); knowledge organization (science-oriented, economics, and administration oriented classification systems; information-systems oriented classification systems).

This framework helps indicate the content to be classified and structural and functional requirements for different kinds of scheme. For instance, classification of "entities", "objects" or "persons" that may be used for various types of scientific or administrative knowledge organization purposes, are most likely to have a simpler structure than classifications that will be created for the organization of literature about these "entities", "objects" or "persons". Classifications created to mediate recorded knowledge, i.e. library and documentary classifications, ought to reflect the multifaceted nature of the way the knowledge is recorded and communicated:

– Aspects of presentation: point of view of the discipline in which the subject is treated.

– Forms in which the knowledge is presented: analytical, historical, critical etc.

– Aspects of the audience: for whom, for what purpose is the document created.

– Forms in which the knowledge is manifested: book, article, study, speech etc.

– Author's points of view.

– Types of document carrier: text, sound, image etc.

In addition, subjects and scientific phenomena often interact and the nature of these interactions and relationships may be the content of a document. Hence, documentary classifications will not only need to express the interaction of any two subjects but also the actual nature of this interaction as this particular aspect may be relevant in information discovery.

3. Limitations in classification expertise

Due to a long history of use, we have a large number of library and documentary classifications. Often, professional expertise in classification is limited to a single scheme and this is especially so with internationally used universal classification systems such as *Dewey Decimal Classification* (*DDC*), *Universal Decimal Classification* (*UDC*), *Library of Congress Classification* (*LCC*), *Bliss Bibliographic Classification* (*BC2*) and *Colon Classification* (*CC*). Once embedded in practice, these schemes are likely to remain the only system of choice in a certain region or type of library for an indefinite period of time. This results in a confined and single-scheme orientated field of training, expertise and research and reflects on the accompanying literature. Hence, we may have subject specialists and consequently authors very familiar with, for instance, *DDC* who may not necessarily know much about *LCC* or even less so about *UDC* or *BC2*.

The exchange, cross-fertilisation or build-up of knowledge on documentary classification is further impeded by the fact that individual classification systems use scheme-specific terminology for what may be common structural or functional features, thus creating a scheme-specific "jargon". In addition, natural competition between classification schemes and between classifications and other indexing languages has also contributed to the narrowing and weakening of professional expertise.

A predominantly specialized and limited knowledge of classification systems in the bibliographic domain have had, for instance, negative consequences for building tools and standards to support the use of classifications in library systems. So although the biggest schemes have already been automated and maintained in databases since the 1980s and 1990s (*DDC*, *UDC*, *LCC*) -it has taken some time to initiate a standard authority data format for use and exchange of classifications in the whole of the bibliographic domain (cf. **Markey**, 2006). At this point, a lack of a common data

model for documentary classifications has resulted in poor and inadequate Marc formats for classification data (**Slavic**; **Cordeiro**, 2004a). The first created was *Marc 21 Concise Format for Classification Data* (cf. 7th update 2006). This standard was based on requirements for online management of *DDC* and *LCC* primarily and was meant to be used for any similar enumerative classification. The *Concise Unimarc Classification Format* followed in 2001 to mirror exactly the same data structure and browsing and retrieval functionality. In their current versions, these formats store the classification notation as a text string and are unable to support schemes with notations that are structured with semantically meaningful elements that need to be searched, browsed and centrally managed. This means that it is not possible to exploit the advantages in knowledge organization and information retrieval that are supported by analytico-synthetic or faceted systems such as *UDC*, *BC2* or *CC*, or for that matter any other faceted classification. For instance, this format does not allow one to code, access, manage and search each part of the composite notation such as, for instance *UDC* number for *94(460)"15" History-Spain-16th century*. As a consequence, libraries using *UDC* classification through a classification authority file will have problems accessing the structural elements of a complex classification number in order to improve subject access in their opacs.

This example illustrates how important it may be that in creating more generally applicable networked standards for controlled vocabularies, we become more informed about documentary classifications in general.

4. Observations on typology of documentary classifications

In this section we are going to highlight some aspects of classification systems and provide some observations that may be relevant for assessing and comparing documentary classifications.

4.1. According to the subject coverage and application area

With respect to subject coverage, classifications are either special or universal. Library and bibliographic services and library networks often develop their own classification system. With respect to this we may have special and universal systems developed and used locally as in-house (or home-grown) schemes. Then we may have schemes developed and used on a national or regional level. And finally, we will find both special and universal systems that are widely spread and used internationally. The history of documentary classifications holds many examples of classifications initially developed for a single library that later became internationally accepted systems. Such is the case of, for example, *LCC* (universal) or the *National Library of Medicine (NLM) Classification* (special). Equally, some systems devised to be used nationally became of interest to other countries. Typology according to subject and usage is well exemplified in the list of "Controlled vocabularies, thesauri and classification systems available in the www" compiled by **T. Koch** (1998).

Irrespective of the subject coverage or application field, documentary classifications may contain from a few hundred to a hundreds of thousands of classes. When we are dealing with classification from the same subject and usage category, it is worth noting that indexing power is not directly proportional to the size of a vocabulary but rather to its quality and the availability of synthetic functions. In order to achieve the same level of accuracy and specificity in indexing, classification that allows the synthesis of complex subject statements needs a much smaller vocabulary than an enumerative system.

4.2. According to the purpose in document and information organization and management

Until now, we have used the term "documentary" as a generic term for both library and bibliographic classifications. At this point we need to make a distinction between the different purposes for which we classify documents, as the original intention behind a classification may determine its structure, syntax, notational system and function in information retrieval. More detailed indexing will, obviously, require a more complex syntax and more rigorous structural patterns.

When classifications are designed to support a systematic, physical arrangement of documents on library shelves primarily, we speak of library classifications proper. The expression bibliographic may be reserved for classifications that are designed for detailed indexing of a great variety of more complex contents such as articles in journals, research studies and reports or non-book materials. Their primary purpose is one of information retrieval (browsing and searching) based on document surrogates in bibliographies or bibliographical databases.

The main purpose of a library classification is "mark and park", i.e. the establishing of a single and the most useful shelf place for a given document. Since notational symbols representing subjects have to be written on book spines, these systems are very concerned with the length of notation and are in favour of a short and compressed notational system.

Because of the fact that physical documents can be assigned only one shelf place these systems will contain more elaborate case-sensitive instructions on how to classify documents dealing with ambiguous con-

tent or multiple subjects. As it happens, classification arranges one subject area at a time and some library schemes may be less concerned with related and associated concepts that are placed in some other subject area. As this will have no impact on the efficiency of classification in supporting systematic shelf browsing, semantic linking and references across distant knowledge areas may happen to be very weak. The advantage of this, however, may be that entire disciplines or subject areas may be "taken out" and used independently as special subject classifications (e.g. *LCC* or *BC2*).

Widely known and used universal library classifications proper are *DDC*, *LCC*, *CC* and *BC2*. There are a large number of special classifications of this type, e.g. *NLM Classification and Mathematical Subject Classification*.

For bibliographic classifications the most important requirements are specificity and accuracy in indexing and the possibility of combining and pre-coordinating subjects. These systems are likely to require more flexibility in connecting and relating different areas of knowledge that may interact in the literature. When universal in coverage, these systems will tend to provide both a sufficient amount of lateral, associative linking between subject areas and the ability to trace the same concept in various disciplines and subject combinations. As they are used primarily with document surrogates they are designed in such a way that each document can be assigned as many meaningful notational elements as necessary to describe the content. Typical examples of classifications primarily built for this purpose are *UDC* and to a certain extent the *Soviet Library Bibliographic Classification*. *The Global Classification for Forestry, Iconclass and Inspec Classifications* are good examples of special classification systems.

Once created, library and documentary classifications are often used interchangeably. When a library classification is used in bibliographic databases one can expect complaints that the system is too limited, rigid and coarse and does not perform well in information retrieval. For example, we may imagine that this would be an objection for a typical library classification such as *DDC* when it is applied to a classification of scientific or research papers. When a typical *bibliographic* classification is used for the arrangement of shelves we can expect complaints about the system being too complex, ambiguous or too detailed. Here a good example is when librarians have to choose how much detail from *UDC* may be necessary for library shelf arrangement. While it may be easy to simplify and downsize a more complex and flexible system, it may prove much harder to add more indexing power for a system that was not designed for detailed indexing in the first place. But if we are aware of the source of the problem we are in a position to adapt the system by adding on or taking off certain structural or notational features.

4.3. According to knowledge organization (macrostructure)

There are three types of knowledge organization structures that are relevant in knowledge mediation: taxonomic, aspect -i.e. disciplinary-based, and phenomena-based. In the first case the expression "taxonomy" is used for the systematic organization of objects/entities classified according to the one, essential principle of division resulting in a taxonomic structure in which each entity appears only once. Taxonomic classifications such as the taxonomy of plants, taxonomy of animals, taxonomy of planets, chemical elements etc. are very typical for knowledge organization in science. In the periodical classification of chemical elements, for example, the one single principle of division "the total number of protons in the atomic nucleus" is applied to all members of the system and in the resulting classification each chemical element will be listed only once in the table.

When knowledge disciplines and sub-disciplines are the primary principle of organizing knowledge, phenomena and associated entities and processes will be subsumed to the aspect of discipline, and we are then dealing with a *perspective* or aspect classification system. In this kind of classification a single phenomenon will appear in any discipline or field of knowledge in which it may be the subject of study. For instance, the concept of "fish" will be listed in the subdivision of zoology, sport, and agriculture.

When the primary principle of organization of knowledge is phenomena, i.e. when the knowledge structure lists phenomena followed by aspects/disciplines of their treatment, such a scheme is called a classification of phenomena. In such a classification, for instance, "fish" would be "the main class" and subclasses would be zoology, agriculture, sport.

In information and documentation, however, we do not deal with entities or phenomena as such but rather with the literature about them. Hence, the same concepts or phenomena may be studied in many fields of knowledge. A "fish" can be analysed in zoology, animal husbandry, the food industry, sport or cooking. Document indexing aims to group similar contents in the way books are likely to be sought and it is assumed that, for instance, a nutritionist looking for "fish cooking recipes" will not necessarily be interested in books on fishing, growing fish or the sport of fishing. Collecting all books about "fish" on a single library shelf does not seem to make sense in practice. Hence, although in the past there were libraries organized according to a classification of phenomena (for example, *Subject*

Classification of Brown) this principle of organization is recognized as ill-suited for library users (**Ranganathan**, 1961). Thus, nowadays most of the widely used documentary classifications are disciplinary, i.e. aspect classifications.

It is worth noting that any disciplinary classification also includes relevant scientific taxonomies and places them within disciplines in which they are the subject of study. For instance, a class of biology in *DDC* or *UDC* contains the taxonomy of plants and animals and these taxonomies are also re-used in the class of agriculture. Most importantly, every documentary classification system will link and reference the same phenomenon or concept across the entire system of knowledge thus enabling, if required, the collocation of information on a single phenomenon irrespective of the field of study. This feature of a system is known as the syndetic structure. Linking of, so called, distributed relatives across the universe of knowledge is one of the features of documentary classifications that can be exploited in information retrieval. Subject-alphabetical indexes to classification, such as the relative-index in DDC is an excellent example of how distributed concepts scattered across the system may actually be linked.

The number of disciplines and main classes of knowledge and their sequence and rigidity which becomes obsolete with the development of science and the emerging of new knowledge has been the subject of much criticism and was extensively written about. It is widely accepted that the disciplinary structure of decimal classifications such as *DDC* and *UDC*, with ten main classes, are very poorly equipped to properly represent the universe of knowledge. Classifications with a wider disciplinary base and logical sequence of disciplines, such as *BC2* which is based on the theory of integrative levels, are in a much better position to provide more appropriate knowledge presentation (see more in: **Beghtol**, 1998; **Broughton**, 2004; **Gnoli**, 2007). It is worth acknowledging, however, that the rigidity of a disciplinary structure can be alleviated if schemes are structured in such a way that they allow the free and flexible combination of simple concepts within and between disciplines.

4.4. According to class/concept organization (microstructure)

The way classes and concepts are organized and coupled with syntactic rules for their combination will determine the power of a classification system in indexing and information retrieval. With respect to this feature we usually make a distinction between enumerative and analytico-synthetic classifications, while in reality systems may combine elements of both approaches. In principle, enumerative classifications belong to the old "library-shelf" type of schemes while analytico-synthetic schemes are more modern systems and better suited for information retrieval purposes.

Enumerative classification schemes "enumerate", i.e. represent in the same manner and in the same hierarchical sequence, all classes irrespective of whether they represent simple or complex concept combinations. When using an enumerative classification one can use classes only in the way they are already listed and there are no means of expressing combinations of two distinct subjects or establishing the relationships between them by connecting or coordinating notational symbols. As a consequence, if a document covers several subjects, one has to choose under which subject to classify the book and has no means of expressing the other equally relevant part of the content. To make up for the lack of synthetic features these schemes try to predict which combination of concepts are likely to be necessary for library shelves and try to list as many combinations as possible.

As opposed to a list of hierarchies in enumerative systems, analytico-synthetic classifications are structured on the principle of mutually exclusive facets of simple concepts that are meant to be combined in the process of indexing. Notational symbols assigned to this structure may contain facet indicators that will allow that these concepts are then synthesised and can be easily split in the process of retrieval (figure 1).

Most analytico-synthetic classifications will keep general concepts that are applicable to all disciplines (places, times, forms, persons, materials, languages etc.) as separate tables. These are usually called auxiliary schedules, auxiliary tables or common isolates, and concepts from these classes can be combined with classes in any field of knowledge. There are analytico-synthetic classifications, however, in which disciplines themselves are organized in a way that allows logical composition and decomposition of classes -and they are called faceted classifications, or more precisely *a faceted classification proper*. In creating a classification structure of a field of knowledge, faceted classifications apply the method of facet analysis based on fundamental facet categories such as entities, kinds, parts, processes, materials or, as **Ranganathan** proposed, personality, matter, energy, space, time. Most widely known representatives of these systems are *Colon Classification* and *Bliss Bibliographic Classification*.

Irrespective of the nature and extent of facet analysis applied, there are several levels of synthesis that may be put in a function in an analytico-synthetic classification (cf. **Gnoli**, 2007; **Slavic**; **Cordeiro**, 2004a; **Isaac**; **Slavic**, 2006):

– Combinations within the same subject field when simple concepts from various facets (entities, proc-

Schedules of an enumerative scheme	Schedules of a faceted scheme
1 Judaism 11 Prayer in Judaism 12 Marriage and the family in Judaism 15 Priests in Judaism 151 Marriage of priests in Judaism 1511 Marriage of priests in Judaism in Europe 2 Christianty 21 Prayer in Christianity 22 Marriage and the family in Christianity 25 Priests in Christianity 251 Marriage of the priests in Christianity 2511 Marriage of the priests in Chr. in Europe ... 3 Islam 31 Prayer in Islam 31 Marriage and the family in Islam 35 Priests in Islam 351 Marriage of priests in Islam 3511 Marriage of priests in Islam in Europe	Places (1) Europe Religion -1 Worship. Prayer -2 Marriage and family -3 Officers of religion. Priests A Judaism B Christianity C Islam

Examples of combination in indexing:

A-1 Prayer in Judaism
A-2 Marriage and family in Judaism
A-3 Priests in Judaism
A-3-1 Marriage of priests in Judaism
A-3-1(1) Marriage of Jewish priests in Europe
...
B-3(1) Christian priests in Europe
B-3-1(1) Marriage of Christian priests in Europe
...
C-3-1(1) Marriage of Islamic priests in Europe
...
A-1:B-1:C-1 Relationships between prayer in Judaism, Christianity and Islam

Figure 1: Class structure in an enumerative and faceted schemes

esses, materials, products) are synthesised to express a combined subject. For instance, within religion we may combine the processes of worship with objects of worship and places of worship

– Combination between subject specific classes/ concepts within a certain subject field with generally applicable concepts or so called common isolates of place, time, persons. For instance between the class of road transport regulations and a class of common concepts of place such as Europe or Spain in, for example, the 20th century.

– Combination between two "distant" main subjects such as ethnography and cinema, or economy and transport, or marketing and agriculture, and expressing the type phase relationships between subjects (influence, comparison, bias, application etc.)

Analytico-synthetic features can have great value in information retrieval as they allow post-coordinate searching and greater flexibility in aggregation and presentation of subjects on the interface using facet-based views. Their usability in browsing and searching will, however, depend on both the notational system in place and the machine readability of a synthesised index.

4.5. According to the notational system

In principle a notational system is an independent element of the scheme which becomes attached to the after the structure itself has been created. At the point of use, each notation is used as an "index term", thus putting the notation into the focus of user concerns. Normally, scheme designers will choose symbols to be numerical, alphabetical or alphanumerical on a completely arbitrary basis. They also have to decide whether to use notation as decimal or as ordinal. It is worth noting that the choice of symbols and their ordering nature is guided primarily by the purpose of the classification (for library shelves or for information retrieval) and not by its structure (enumerative or analytico-synthetic).

Notation can be expressive of hierarchy and of syntax or of both. Hierarchically expressive notation mirrors the hierarchy and each digit or letter of the notation will represent one level in division. The deeper in the hierarchy the concept is, the longer the notation. Clas-

sifications with hierarchically expressive notations are much friendlier to navigate and use. When presenting a classification hierarchy in print or online there will be no need to show the indentation of subordinate classes as this will be obvious from the notation itself, e.g.:

1 Mammalia. Mammals
11 Carnivora. Carnivorous mammals
111 Fissipedia. Terrestrial carnivores
1111 Ursidae. Bears
11111 Polar bear
11112 Brown bear

When notations have expressive syntax this makes it easy to compose and decompose the notation. In such a system, a notation from one facet hierarchy will always have the same beginning, so called facet indicators (symbols, punctuations, letters or numbers), which will keep notational elements in a synthesised expression easily distinguishable and easy to link to their verbal expressions:

11112 Brown bears
-51 Metabolism
-A Hibernation
synthesized classmark

11112 -51 – ABrown bear – Metabolism – Hibernation

Expressive notations tend to be longer and more complicated and are usually considered less suitable for shelf arrangement and labelling of books. They are however more appropriate for use in an online environment where the rich and informative notation can have advantages in supporting data for searching, browsing or classification presentation. A good example of a classification with notation that is expressive with respect to hierarchy is *DDC*, while *CC* and *UDC* are examples of notations which are, to a great extent, both hierarchically and syntactically expressive. An example of a fully expressive notation can be seen in a recently created special classification *FAT-HUM* (**Broughton**; **Slavic**, 2007).

Non-expressive notation is usually chosen when there is the need to have very short symbols for book labels in a systematic shelf arrangement. Good examples are *LCC*, *BC2* and *BSO*. In classifications with this type of notation, presentation of the schedule in print and online requires greater effort as the only information on a hierarchy is the indentation of class description e.g.:

11 Insecta (Hexapoda). Insects. Entomology
12 Orthoptera
13 Dermaptera. Forficulidae. Earwigs

It is has been already acknowledged that formats for expressing classifications in a machine readable way are expected to be sufficiently supportive of all semantic relationships (hierarchy/associative relationships), structural features and syntactic rules that we were observing here (**Gödert**, 1991; **Pollitt**; **Tinker**, 2000; **Slavic**; **Cordeiro**, 2004). When this is the case, browsing and searching of a classification is not based on a notation only but rather on the underlying data structure recorded in a classification authority file (**Slavic**; **Cordeiro**, 2004a). But most importantly, with the help of authority control, searching can be performed using words, as classification numbers may be linked to natural language terms in one or more languages (**Pika**, 2007). This means that one is able to move/expand from one subject to a broader subject area without even seeing a classification number. When a classification has a fully expressive notation it takes less effort to analyse and control the structure and convert it to a machine readable format or process it for the purpose of searching (e.g. automatic decomposition of notation). For instance, the subject of "History of Spain in 16th century", expressed in *UDC* as *94(460)"15"* can be retrieved either by looking for *(460)* or *"15"* or a combination of both using boolean logic. Searches can be expanded by looking for the broader category of *(460)* Spain which is *(46) Iberian peninsula* or even broader to *(4) Europe*. Logically, if a scheme does not have an expressive notation similar to the one illustrated in the example of *UDC*, its automation requires additional input from classification specialists and for large systems this can be painstakingly slow and expensive. Also, the attaching of natural language terms to classification notation may also be more difficult (**Slavic**; **Cordeiro**, 2004).

We have attempted to demonstrate here that knowledge of some of the intricate and technical details of different documentary classification systems can help understand their potentials and limitations but can also help when creating standards that are aimed at their exploitation and use in a networked environment.

5. Concluding remarks

In any advanced information seeking scenario based on a controlled vocabulary, the vocabulary itself is meant to be managed centrally and independently from the metadata or resources themselves. This can be achieved by following the bibliographic model of subject authority control but also by adopting the solution of terminology services. In either case classifications and other controlled vocabularies would need to be expressed in a machine readable format that would allow the exploitation of semantic relationships between concepts. With the available technology we are in a position to harvest and exploit existing subject data. Collections indexed by documentary classifications represent a wealth of semantically organized information which

we can render more useful providing we learn more about the systems by which they are organized.

Understanding to which level and in which way hierarchy or syntax relationships are expressed in a certain scheme may help in unlocking and capturing the semantic relationships that may be otherwise lost. For instance, if we do not like a notational system of an otherwise satisfactory scheme-scheme management online will allow us to "detach" the notational system, hide it or replace it with a system of symbols that is more suited to our purpose without losing any of the functionality or interoperability of the system itself. Or, if we do not like the disciplinary structure of a system we may build facets or interdisciplinary orientated views using and building further on associative and lateral relationships within the system. As it becomes easier to build, maintain and implement or change classifications it becomes more important to share and disseminate knowledge on the technicalities of building and using classifications that will make their application cheaper and more efficient.

Bibliography

Aitchison, J. "The thesaurofacet: a multipurpose retrieval language tool". En: *Journal of documentation*, 1970, v. 26, n. 3, pp. 187-203.

Beghtol, C. "General classification systems: structural principles for multidisciplinary specification". En: **Mustafa Elhadi, W.**; **Maniez, J.**; **Pollitt, S**. (eds.). *Structures and relations in knowledge organization: proceedings of the Fifth International ISKO Conference*. Würzburg: Ergon Verlag, 1998, pp. 89-96.

Bhattacharyya, G. "Fundamentals of subject indexing languages". En: **Neelameghan, A.** (ed.). *Ordering systems for global information networks: proceedings of the Third international study conference on classification research*. Bangalore: DRTC: FID/CR and Sarada Ranganathan Endowment for Library Science, 1979, pp. 83-99.

Bhattacharyya, G. "Classaurus: its fundamentals, design and use". En: **Dahlberg, I.** (ed.). *Universal classification: subject analysis and ordering systems: proceeding of the 4th International study conference on classification research, vol. 1*. Frankfurt: Indeks Verlag, 1982, pp. 139-148.

Broughton, V. *Essential classification*. London: Facet Publishing, 2004.

Broughton, V.; **Slavic, A.** "Building a faceted classification for the humanities: principles and procedures". En: *Journal of documentation*, 2007, v. 63, n. 5, pp. 727-754.
http://dlist.sir.arizona.edu/1976/

BS 8723-1; BS 8723-2. Structured vocabularies for information retrieval. Part 1: definitions, symbols and abbreviation; Part 2: thesauri. London, British Standards Institution, 2005.

Concise Unimarc classification format (draft), 2001.
http://www.ifla.org/VI/3/p1996-1/concise.htm

Dahlberg, I. "The basis of a new universal classification system seen from a philosophy of science point of view". En: **Williamson, N. J.**; **Hudon, M**. (eds.). *Classification research for knowledge representation and organization: proceedings of the 5th International study conference on classification research*. Amsterdam: Elsevier Science Publishers; The Hague: FID, 1992, pp. 187-209.

Desire information gateway handbook, 2000.
http://www.desire.org/handbook/2-5.html

Foskett, A. C. *The subject approach to information*. 5th ed. London: Library Association Publishing, 1997.

Fugmann, R. "Unused opportunities in indexing and classification". En: **Fugmann, R.** (ed.). *Tools for knowledge organization and the human interface: proceedings of the 1st International ISKO conference.* Frankfurt/Main: Indeks Verlag, 1990, pp. 65-77.

Gnoli, C. "Progress in synthetic classification: towards unique definition of concepts". En: *Extensions & corrections to the UDC*, 2007, n. 29 [in print].
http://dlist.sir.arizona.edu/1945/01/synthetic.pdf

Gödert, W. "Facet classification in online retrieval". En: *International classification*, 1991, v. 18, n. 2, pp. 98-105.

Hodge, G. *Systems of knowledge organization for digital libraries*, 2000.
http://www.clir.org/pubs/reports/pub91/contents.html

Isaac, A.; **Slavic, A.** *UDC: the Universal Decimal Classification*.
http://www.w3.org/2006/07/SWD/wiki/EucUDC

ISO 5963. Documentation-methods for examining documents, determining their subjects, and selecting indexing terms: international standard. Geneva, International Organization for Standardization, 1985.

ISO/IEC 13250:2000 Topic maps: information technology-document description and markup languages. **Biezunski, M.**; **Bryan, M.** (eds.). Newcomb. First, December 3 1999.
http://www.y12.doe.gov/sgml/sc34/document/0129.pdf

Koch, T.; **Day, M.** *The role of classification schemes in internet resource description and discovery: Desire project report*. Bath: Ukoln, 1997.
http://www.ukoln.ac.uk/metadata/desire/classification/

Marc 21 concise format for classification data. Update no. 7, 2006.
http://www.loc.gov/marc/classification/

Markey, K. "Forty years of classification online: final chapter of future unlimited?". En: **Mitchell, J. S.**; **Vizine-Goetz, D**. *Moving beyond the presentation layer: content and context in the Dewey Decimal Classification (DDC) System*. Binghampton, NY: The Haworth Information Press, 2006, pp. 1-63.

Pika, J. "Universal Decimal Classification at the ETH-Bibliothek Zurich-a Swiss perspective". En: *Extensions & corrections to the UDC*, 2007, n. 29 [in print].

Pollitt, S.; **Tinker, A. J.** "Enhanced view-based searching through the decomposition of Dewey Decimal Classification Codes". En: **Beghtol, C.**; **Howarth, L. C.**; **Williamson, N. J.** (eds.). *Dynamism and stability of knowledge organization: proceedings of the Sixth International ISKO conference*. Würzburg: Ergon Verlag, 2000, 288-294.

Ranganathan, S. R. "Library classification on the march". En: **Foskett, D. J.**; **Palmer, B. I.** (eds.). *The Sayers memorial volume: essays in librarianship in memory of William Charles Berwick Sayers*. London: The Library Association, 1961, pp. 72-95.

Slavic, A.; **Cordeiro, M. I.** "Core requirements for automation of analytico-synthetic classifications". En: **McIlwaine, I. C**. *Knowledge organization and the global information society: proceedings of the 8th ISKO conference*. Würzburg: Ergon Verlag, 2004, pp. 187-192.

Slavic, A.; **Cordeiro, M. I.** "Sharing and re-use of classification systems: the need for a common data model". En: *Signum*, 2004a, n. 8, pp. 19-24.

Svenonius, E. "Precoordination or not?". En: **Holley, Robert P.**, et al (eds.). *Subject indexing: principles and practices in the 90'*. München: K. G. Saur, 1995, pp. 231-255.

Svenonius, E. *The intellectual foundation of information organization*. Cambridge, MA; London: The MIT Press, 2000.

Tudhope, D.; **Koch, T.**; **Heery, R.** *Terminology services and technology: JISC state of the art review*, 2006.
http://www.ukoln.ac.uk/terminology/JISC-review2006.html

Vizine-Goetz, D., et. al. "Vocabulary mapping for terminology services". En: *Journal of digital information*, 2004, v. 4, n. 4.
http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Vizine-Goetz/

***Aida Slavic**, UDC Editorial Team, UDC Consortium, PO Box 90407, 509 LK The Hague, The Netherlands*
aida@acorweb.net

LIBRERÍAS TROA






Innovación y diseño de servicios culturales para la formación de la persona

A CORUÑA
AVIR Tel. 981 27 31 17
c/Juan Flórez, 30.

BARCELONA
GARBÍ Tel. 932 17 54 08
Via Augusta, 9.
GARBÍ IESE Tel. 932 53 42 00
Av. Pearson, 21.
GARBÍ UIC Tel. 934 17 59 30
c/Inmaculada, 22.
GARBÍ SAN CUGAT Tel. 935 04 20 25
c/Garriera, s/n.

BILBAO
OLERKI Tel. 944 23 57 55
c/Marqués del Puerto, 1

CÁCERES
BUJACO Tel. 927 22 20 19
Av. Virgen de la Montaña, 2.

GIRONA
EMPÚRIES Tel. 972 20 34 29
c/Álvarez de Castro, 6.

GRANADA
DAURO Tel. 958 22 45 21
c/Zacatín, 3.

MADRID
NEBLÍ Tel. 915 76 21 03
c/Serrano, 80.
NEBLÍ IESE
Camino del Cerro del Águila, 3
Ctra. Castilla s/n.
DELSA Tel. 914 35 74 21
c/Serrano, 80.
PROXIMAMENTE ARAVACA-POZUELO
Av. Europa 25
PROXIMAMENTE LAS TABLAS
Paseo Tierra de Melide 13.

MÁLAGA
JÁBEGA Tel. 952 22 29 23

PAMPLONA
LIBRERÍA UNIVERSITARIA Tel. 948 17 02 90
Av. Sancho el Fuerte, 24
TIENDA UNIVERSITARIA Tel. 948 26 72 25
Nuevo Edificio de Bibliotecas.
UN Campus Universitario.
CLÍNICA UNIVERSITARIA Tel. 948 25 54 00
Av. Pío XII, 36.

SAN SEBASTIÁN
ZUBIETA Tel. 943 42 70 08
Plaza de Guipúzcoa, 11.

SEVILLA
TARSIS Tel. 954 21 25 65
c/Luis de Morales, 1.
SAN TELMO Tel. 954 97 50 04
Av. Mujer Trabajadora, 1.

VALENCIA
IDEAS Tel. 963 34 83 18
c/Grabador Esteve, 33.

ZARAGOZA
FONTIBRE Tel. 976 21 53 96
c/Canfranc, 9.

A CORUÑA BARCELONA-GARBÍ BARCELONA-IESE BARCELONA-UIC BARCELONA-SAN CUGAT BILBAO CÁCERES GIRONA GRANADA
MADRID-DELSA MADRID-NEBLÍ MADRID-IESE MADRID-POZUELO MADRID-LAS TABLAS MÁLAGA PAMPLONA-UNIVERSITARIA
PAMPLONA-TIENDA UNIVERSITARIA PAMPLONA-CLÍNICA UNIVERSITARIA SAN SEBASTIÁN SEVILLA-TARSIS SEVILLA-SAN TELMO
VALENCIA ZARAGOZA

Metodología para la estructuración del conocimiento de una disciplina: el caso de *PuertoTerm*[1]

Por **Jose A. Senso, Pedro-Javier Magaña-Redondo, Pamela Faber-Benitez** y **Amparo Vila-Miranda**

Jose A. Senso *es doctor en documentación y profesor del Departamento de Biblioteconomía y Documentación de la Universidad de Granada. Imparte su docencia en la Facultad de Traducción e Interpretación donde además es vicedecano de nuevas tecnologías. Es miembro del grupo de investigación Lexicon y de la Red Temática de Documentación.*

Pedro Javier Magaña Redondo *es ingeniero informático por la Universidad de Granada, y diplomado en estudios avanzados en diseño, análisis y aplicaciones de sistemas inteligentes. En la actualidad centra su investigación en el estudio y desarrollo de tecnologías que faciliten la adopción de la web semántica por parte de un mayor número de usuarios.*

Pamela Faber-Benitez enseña terminología y lingüística aplicada en la niv. de Granada (UGR), donde es catedrática de traducción e nterpretación. Es licenciada en CC de la información por la Univ. de Carolina del Norte (EUA) y en filología inglesa por la UGR. Realizó su tesis doctoral en lingüística en la Univ. de Paris IV (La Sorbona) y la UGR. Ha dirigido tres proyectos de investigación en gestión de terminología multilingüe en medicina e ingeniería. Es autora de varios libros y artículos sobre semántica léxica, traducción y terminología.

María-Amparo Vila Miranda *es catedrática de universidad desde 1992. Obtuvo el grado de doctor en matemáticas por la Universidad de Granada, donde actualmente dirige el grupo de investigación Idbis (Intelligent Databases and Information System). Es autora de numerosas publicaciones en libros y revistas científicas internacionales, y ha participado como ponente en un gran número de congresos. La mayoría de sus trabajos están relacionados con el diseño de bases de datos, la recuperación de información y datamining.*

Resumen: *En este trabajo se describen los pasos llevados a cabo dentro del proyecto de investigación PuertoTerm para formalizar y estructurar el conocimiento desarrollado dentro de la Ingeniería de Puertos y Costas. Se analiza el proceso seguido para la obtención de la terminología asociada, cómo se ha estructurado y cómo se ha gestionado. Por último, se describe el desarrollo e implementación, prestando especial atención al diseño de la interfaz bidimensional de visualización de la información.*

Palabras clave: *Estructuración del conocimiento, Visualización de la información, Ontologías, Ingeniería de puertos y costas, Modelo lexemático-funcional, Semántica de marcos.*

Title: Methodology for knowledge structuring in a discipline: the PuertoTerm case

Abstract: *The PuertoTerm research project reports the steps taken to formalize and structure the knowledge developed within the field of coastal and ports engineering. The process used to obtain the associated terminology is described and the way in which the terminology was structured and processed is analysed. Finally, development and implementation are described, with special attention to the design of a two-dimensional information visualization interface.*

Keywords: *Knowledge structure, Information visualization, Ontologies, Coastal and ports engineering, Lexematic-functional model, Semantic frameworks.*

Senso, **Jose A.**; **Magaña-Redondo**, **Pedro Javier**; **Faber-Benitez, Pamela**; **Vila-Miranda, Amparo**. "Metodología para la estructuración del conocimiento de una disciplina: el caso de PuertoTerm". En: *El profesional de la información,* 2007, noviembre-diciembre, v. 16, n. 6, pp. 591-604.

DOI: 10.3145/epi.2007.nov.06

Artículo recibido el 29-06-07
Aceptación definitiva: 12-09-07

1. Introducción

El crecimiento de la información en determinadas áreas dentro de la web ha generado un aumento desigual en lo que se refiere a la aparición de herramientas terminológicas que faciliten la recuperación de información (taxonomías, tesauros y ontologías). Así, dependiendo del nivel de asentamiento que tenga una disciplina determinada, será más o menos complejo localizar productos que permitan estructurar el conocimiento asociado a dicha rama del saber.

Un ejemplo claro lo encontramos en la ingeniería de puertos y costas. Aunque es relativamente nueva, la notable evolución científica que ha experimentado debida al desarrollo de nuevas técnicas de construcción, explotación y gestión de estructuras y recursos marinos no ha sido suficiente como para que exista una formalización de las estructuras de conocimiento que representen los procedimientos, definiciones, objetos y metaconocimiento que en ella se llevan a cabo.

Este trabajo refleja el esfuerzo realizado por el proyecto de investigación *PuertoTerm* en la representación de la estructura conceptual del dominio de la ingeniería de puertos y costas, con la finalidad principal de apuntar los posibles pasos para realizar futuras representaciones de la estructura del conocimiento de una disciplina concreta y ayudar de esta manera a la generación de nuevas herramientas para una mejor recuperación de la información.

La mayoría de miembros de dicho proyecto han formado parte de otros dos anteriores que han servido como base teórica y especialmente metodológica para el que se explica a continuación. El objetivo del primer proyecto fue desarrollar una lógica léxica para la traducción asistida por ordenador a partir de una base de datos y el segundo, de gran impacto dentro de la comunidad médica, permitió la elaboración de un sistema de información integrado en internet y dedicado al subdominio de la oncología. Su nombre fue *Oncoterm*.

http://www.ugr.es/~oncoterm

1.1. Objetivos del proyecto *PuertoTerm*

Como ya se ha mencionado anteriormente, la ingeniería de puertos y costas tiene ciertas peculiaridades que la hacen "diferente" al resto de las ingenierías. El hecho de no contar con diccionarios específicos, la ausencia de teoría concreta y la necesidad de los profesionales que trabajan en ella de poseer ambas herramientas para realizar su trabajo, le confiere un cierto halo de área discriminada (al menos hablando desde el punto de vista de la documentación y la terminología). Con *PuertoTerm* se pretendió solucionar esto partiendo del hecho de que para transmitir conocimiento especializado es necesario emplear un sistema que permita especificar el componente semántico inherente a cualquier área del saber.

"Para transmitir conocimiento especializado es necesario emplear un sistema que permita especificar el componente semántico inherente a cualquier área del saber"

El proyecto de investigación se basó en el programa del *Ministerio de Obras Públicas y Urbanismo* "Recomendaciones para obras marítimas", que se inició en 1987 con la constitución de la Comisión Técnica encargada de redactar una serie de recomendaciones para guiar a organismos estatales y empresas privadas en el proyecto, construcción, mantenimiento y explotación de las construcciones marinas. Este programa también dio lugar a la publicación *ROM 0.0* (*Puertos del Estado*, 2001), que proporciona un conjunto de normas y criterios técnicos para la gestión de obras portuarias y marítimas, independientemente de su clase y materiales empleados. La traducción de esta norma al inglés por parte de la investigadora principal del proyecto puso de manifiesto que esta rama de la ingeniería emplea una serie de conceptos nuevos y neologismos con una designación original en español. Esto difiere notablemente de lo que sucede en otros campos de investigación, caso de la medicina por ejemplo, donde el término original procede principalmente del inglés. Por lo tanto, poseer un mayor número de recursos terminológicos en esta área facilitaría la adquisición y transmisión de conocimientos entre expertos de diferentes países.

Los objetivos específicos sobre los que sustentó el proyecto fueron:

– Crear un corpus de textos específicos de esa materia en español, inglés y alemán, así como desarrollar un protocolo de gestión de la información de dichos textos.

– Definir cuáles serían los conceptos y los términos que desarrollan la arquitectura semántica de esa disciplina y establecer las relaciones conceptuales específicas.

– Diseñar y alimentar una base de conocimiento terminológico articulado en torno a la estructura hallada en la definición de los términos. En ella se deberían poder almacenar además las relaciones semánticas existentes entre los términos, de forma que posteriormente se pueda producir una conversión o vinculación a formas de representación más expresivas.

– Crear una aplicación informática que permitiera la recuperación de la información sin que se perdiera la estructura y relaciones formadas entre los conceptos y los términos.

– Crear un banco de imágenes para complementar y enriquecer las representaciones lingüísticas de los conceptos pertenecientes al campo especializado que, por sus características específicas, necesita una representación conceptual más visual y dinámica.

Como se puede observar, la consecución de estos puntos debería plasmarse en la construcción de la estructura conceptual de ingeniería de puertos y costas. De sus términos y conceptos con sus respectivas definiciones, pero también de sus procesos, aplicaciones y herramientas. El presente texto explica los pasos llevados a cabo para proponer una metodología para futuros proyectos similares.

1.2. Fundamentos teóricos

Realizar la estructuración conceptual de un dominio determinado se consigue principalmente por medio de la elaboración de jerarquías terminográficas que se obtienen a partir de la selección y extracción de la información conceptual de textos especializados, diccionarios, tesauros, etc. En realidad, estos principios sirven de base metodológica para varias disciplinas relacionadas con la lingüística (terminología, lexicografía, traducción especializada) y con la documentación (diseño y gestión de tesauros y ontologías).

En el caso de la terminología, desde finales de 1980 se ha generalizado el uso de textos especializados como fuente de información y localización de vocablos técnicos para alimentar bases de datos terminológicas. En la actualidad, y gracias al avance de las tecnologías de la información, estos corpus suelen encontrarse en formato electrónico (**McEnery**; **Wilson**, 1996). La documentación, por su parte, emplea esta misma técnica para la realización de herramientas propias de los lenguajes documentales, como los tesauros y, gracias al desarrollo de conceptos como la web semántica, las ontologías. Una vez más queda patente la estrecha relación que existe entre ambas áreas y que ha sido defendida por muchos autores (**Irazazábal**, 1996; **Pérez Álvarez-Ossorio**, 1988; **Pinto**; **Cordón**, 1999; **Sales**, 2006). Si bien es cierto que la finalidad de ambas ha sido diferente (la terminología establece relaciones entre conceptos, y la documentación se centra en la recuperación de la información), en el proyecto *PuertoTerm* se ha intentado (y creemos que logrado) aunar ambas.

Para llevar a cabo la estructuración del conocimiento de la ingeniería de puertos y costas se han empleado dos teorías lingüísticas con base semántica: el modelo lexemático-funcional de **Martín-Mingorance** (1989; 1995; **Faber**; **Mairal**, 1999) y la semántica de marcos –*frame semantics*- (**Fillmore**, 1982; **Fillmore**; **Atkins**, 1998; **Gahl**, 1998).

El modelo lexemático-funcional facilita la representación de relaciones conceptuales y su posterior

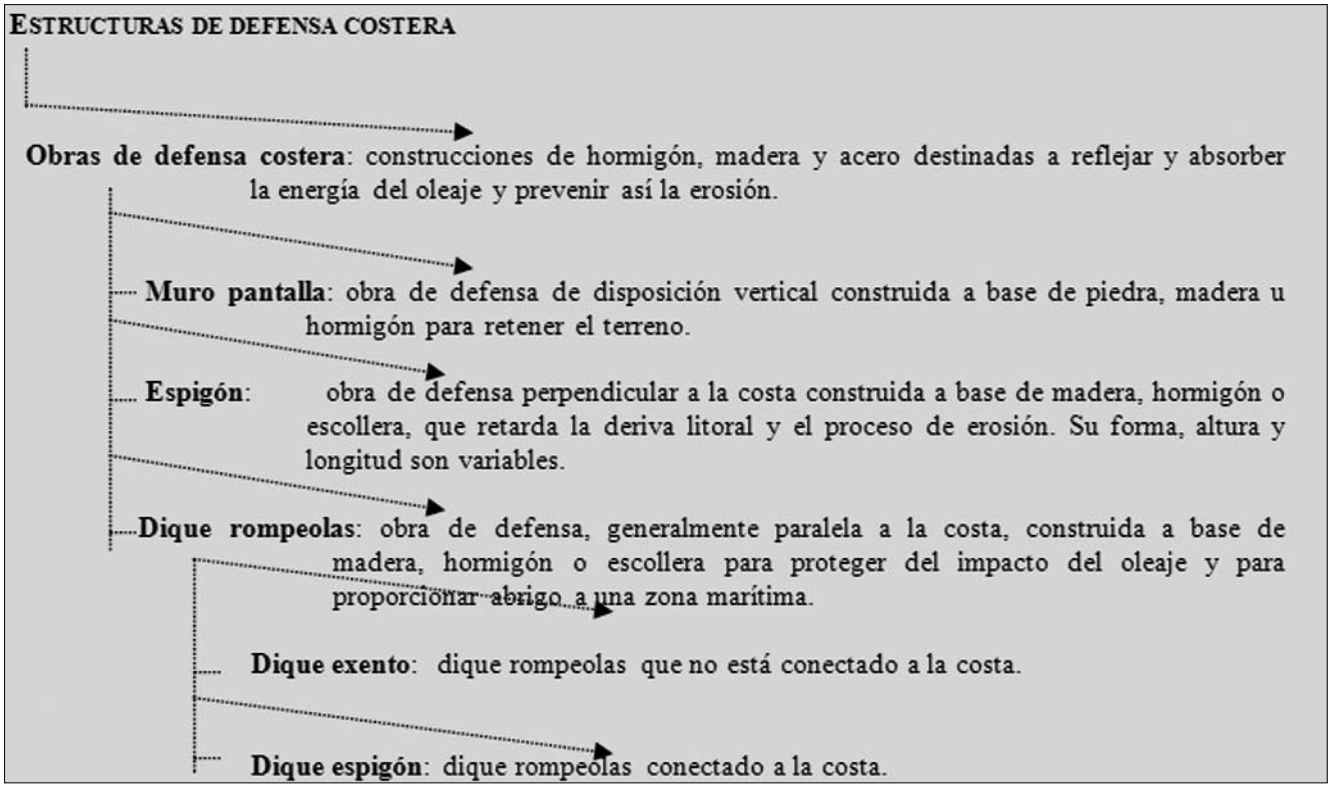


Figura 1. "Estructuras de defensa costera" tiene como subdominio "Obras de defensa costera", que muestra el siguiente diseño, explícito en las definiciones de los términos

clasificación dentro de un lenguaje general y/o especializado. Básicamente propone una distinción entre relaciones sintagmáticas y paradigmáticas –basada en los principios complementarios de combinación y selección (**Lyons**, 1977)- que permite una organización conceptual independiente del sistema lingüístico. Ofrece una metodología lingüística para la organización de conceptos. De hecho, concibe nuestro propio lexicón mental como una compleja red en la que cada nodo es un concepto y estos están interconectados por diferentes tipos de relaciones (**López-Rodríguez**; **Tercedor-Sánchez**; **Faber-Benitez**, 2006). Por ejemplo, en la figura 1 se puede observar la estructuración de "Obras de defensa costera" como subdominio de "Estructuras de defensa costera".

La semántica de marcos (**Fillmore**; **Johnson**; **Petruck**, 2003), por su parte, entiende el concepto de marco como una representación esquemática de un conjunto de conceptos relacionados entre sí. El proyecto *FrameNet*, de la *Universidad de Berkeley* se basa en dicha teoría. Su gran ventaja radica en que la utilización de un único concepto activa todo el sistema conceptual permitiendo, gracias al empleo de enlaces web en su fase de implementación "física", conectar todos los conceptos subordinados y superordinados, facilitando la representación genérico-específica de los términos. Un desarrollo más profundo de esta fase del proyecto se puede encontrar en **Faber** (2006).

http://framenet.icsi.berkeley.edu/

2. Metodología

Para realizar una verdadera estructuración del conocimiento dentro de una disciplina no basta con obtener unas listas de términos y delimitar sus posibles relaciones jerárquicas, ya que son demasiado estáticas y no ofrecen la verdadera visión dinámica que tiene cualquier área del saber. La mayoría de proyectos de este tipo que se basan en lenguajes documentales intentan recopilar términos o plasmar el conocimiento realizando traducciones de los utilizados en los vocabularios de la disciplina que se pretende representar. Esto elimina cierto nivel de comprensión a la hora de entender tareas y procesos, es decir, plasmar algo más que una simple jerarquía. Para evitarlo, el enfoque global que se pretende dar es el que denominamos gestión de terminología orientada a procesos (**Faber** et al., 2005).

La primera fase se inicia empleando técnicas de recuperación de información desarrolladas en el ámbito de la lingüística de corpus, disciplina que ha mostrado su utilidad en proyectos a gran escala como las dos fases de *Acquilex* o, a menor escala dentro del mundo de los lenguajes documentales, los que pretenden generar clasificaciones, tesauros u ontologías (*ISO-5964*, 1985).

http://www.cl.cam.ac.uk/research/nl/acquilex/

Estas técnicas son muy útiles, ya que desde el punto de vista práctico suponen una forma fácil y asequible de obtener gran cantidad de terminología asociada a un área. Desde el punto de vista de uso real de los vocablos, el investigador se asegura que la información extraída es fiel reflejo tanto de los contenidos reales, en nuestro caso de la ingeniería de costas, como del sublenguaje especializado empleado en el mismo. En el caso particular de este proyecto, después de estudiar los recursos terminológicos, se han creado otros nuevos que no existían antes en esta disciplina al ser un dominio muy joven y dinámico y con una terminología en proceso de fijación y estandarización. En otros casos con campos más formados este proceso no tiene por qué ser necesario. Como es habitual, las fuentes de información empleadas para este análisis de corpus han sido obras de referencia especializadas (enciclopedias y diccionarios técnicos mayoritariamente), monografías y artículos de publicaciones científicas. Todas ellas se almacenaron en ficheros de texto en diferentes directorios, dependiendo del idioma en el que estuviese expresado el texto, para facilitar su procesamiento por el resto de herramientas.

"Después de estudiar los recursos terminológicos existentes sobre esta materia, se han creado otros nuevos"

En esta fase, los términos recogidos forman una lista plana, donde ninguno tiene más valor semántico que otro. El siguiente paso se centró precisamente en elaborar un compendio de los conceptos más importantes. Para ello, y con la ayuda de especialistas y de la herramienta de análisis léxico *Wordsmith Tools*, distribuida por *Oxford University Press* y que permite explotar grandes conjuntos de textos mediante búsquedas basadas en parámetros contextuales o estadísticos, se realizó una lista de frecuencias que permitió inferir el conocimiento especializado de la ingeniería de puertos y costas. La identificación de las palabras clave permitió el modelado conceptual de la ontología que articularía todo el conocimiento de esta disciplina. Los lemas más frecuentes permiten identificar las categorías conceptuales sobre las que se fundamenta la definición de los términos del texto. La figura 2 muestra el resultado del análisis del corpus en español realizado por *Wordsmith Tools* a partir de los documentos extraídos tras la primera fase del proyecto.

http://www.lexically.net/wordsmith/

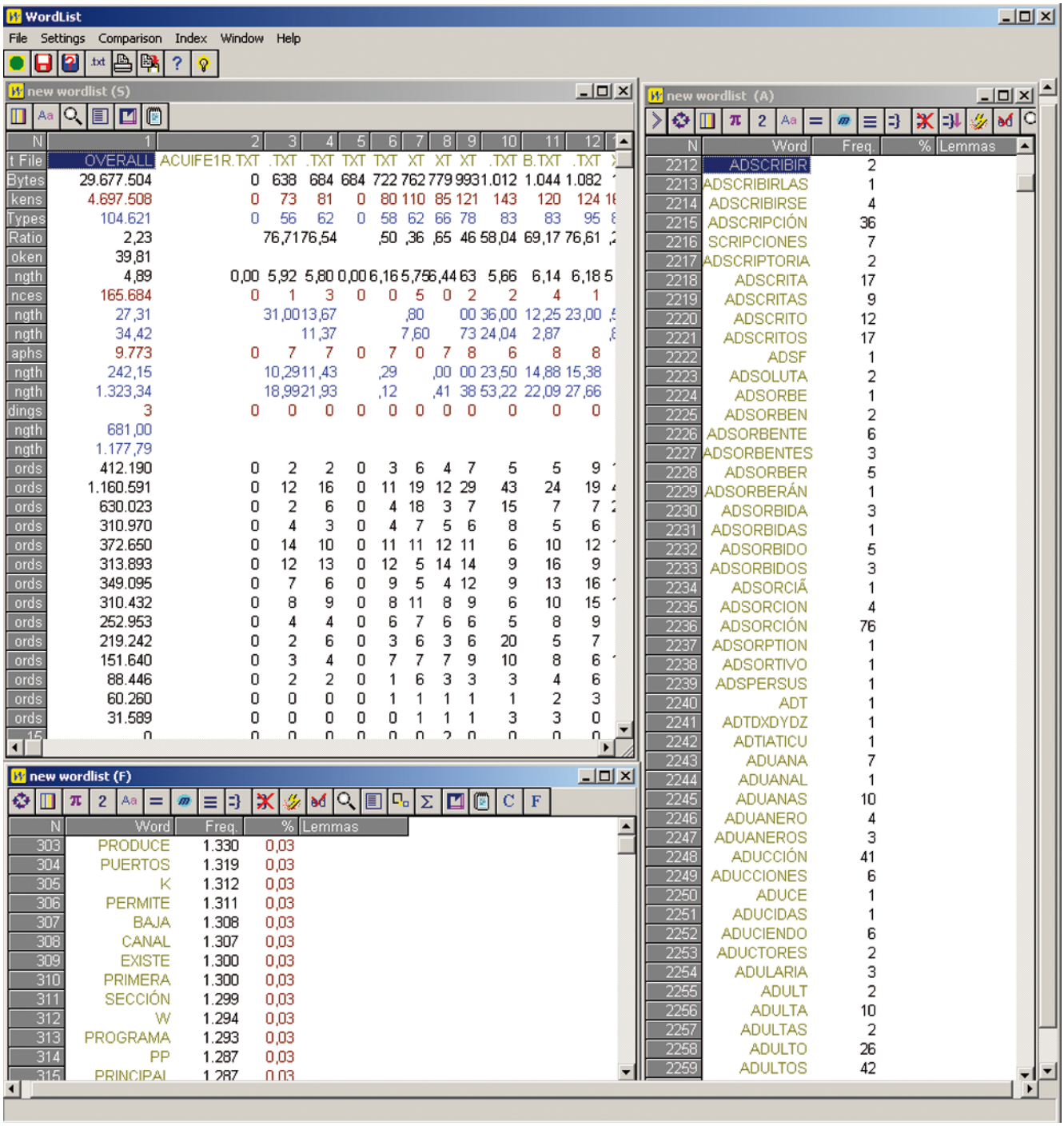


Figura 2. Resultado mostrado por Wordsmith Tools del análisis de los ficheros de texto localizados tras el análisis del corpus. Se puede observar la estadística general (izquierda arriba), la lista de frecuencia de los términos (izquierda abajo) y la lista alfabética de términos (derecha)

Con el fin de conseguir una mayor adecuación en el empleo de cada término, y poder enmarcarlo dentro de su uso correcto, no sólo se utilizaron técnicas estadísticas. Otro elemento clave fue el uso de concordancias, es decir, la presentación de las ocurrencias de una palabra en su contexto lingüístico, de tal forma que la palabra en cuestión aparece en el centro de una línea y a ambos lados se muestran otras con las que suele aparecer en los textos. Se trata de una técnica muy similar a la empleada en los índices *KWIC* (*keyword in context*) pero con una finalidad diferente, ya que aquí se emplea menos para recuperar y más para contextualizar.

Hasta ahora los pasos seguidos nos han permitido adquirir toda la información, en forma de palabras que pueden expresar conceptos o términos, relacionadas con la ingeniería de puertos y costas que nos interesa, y además tenemos cierta consciencia sobre el valor que tienen esas palabras en función de su peso conceptual.

Cada disciplina está caracterizada por poseer uno o varios eventos (o marcos conceptuales) que describen procesos o acciones que tienen lugar dentro de ella. Dichos eventos poseen una estructura que facilita la creación y organización de los diferentes conceptos

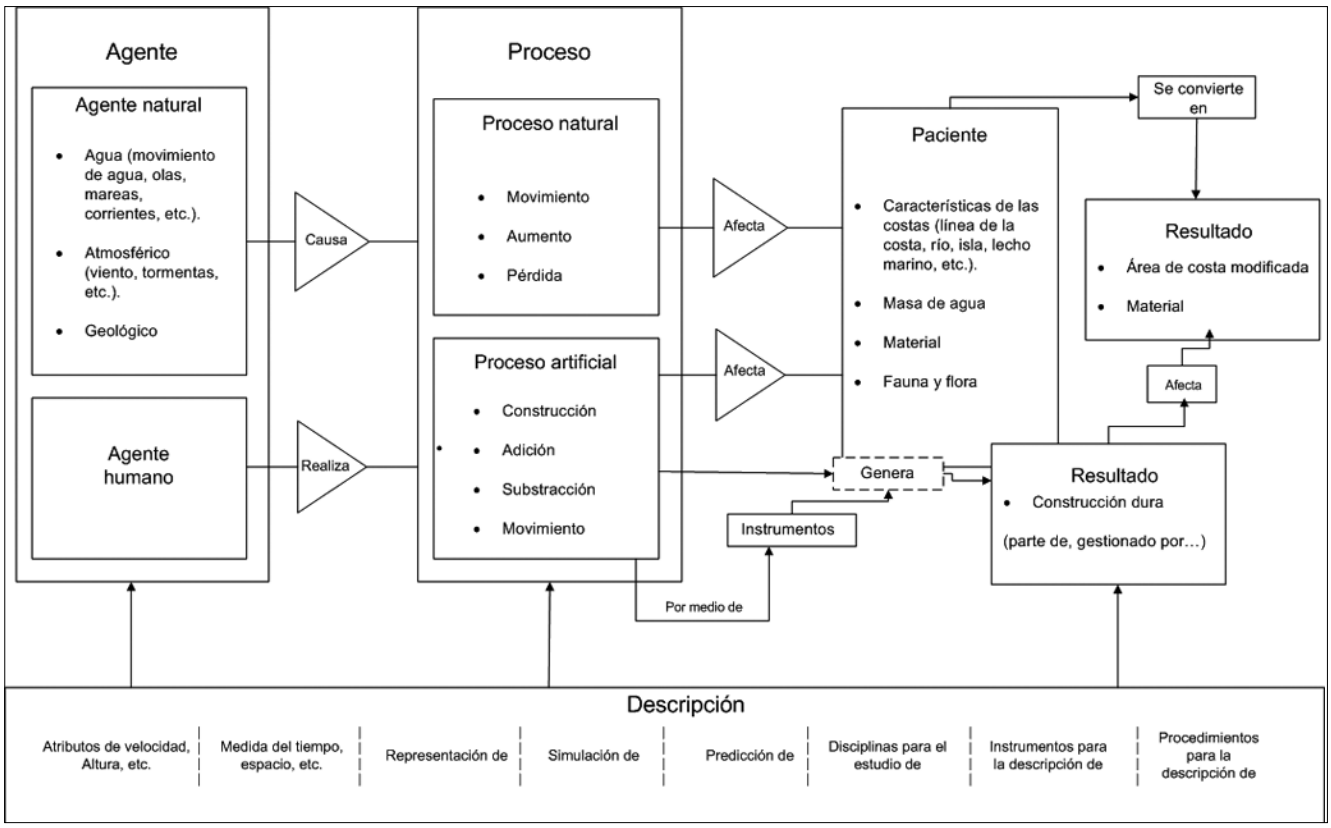


Figura 3. Descripción de los procesos llevados a cabo en la ingeniería de puertos y costas. El análisis del corpus y su posterior categorización permiten conocer cuáles son los eventos que se producen dentro de una disciplina

que los integran. La elaboración del marco conceptual es el culmen del análisis de las definiciones lexicográficas y de corpus extraídas previamente. En nuestro caso, dicho análisis se plasma en la figura 3, que muestra el marco genérico de la ingeniería de puertos y costas.

A partir de ella podemos deducir que todos los procesos llevados a cabo dentro de la ingeniería de puertos y costas son iniciados por un agente que, por medio de una serie de acciones (procesos naturales o artificiales), afecta a un paciente (la costa, la flora y fauna, etc.) y produce un resultado determinado. Estas macrocategorías (agente, proceso y paciente/resultado) constituyen los roles principales del dominio y sirven para proporcionar un modelo que permita representar las relaciones. Como suele ocurrir en muchas áreas, los agentes se pueden valer de instrumentos para generar determinados resultados. De esa forma contemplamos la posibilidad de añadir las herramientas que se suelen emplear en esta materia para, por ejemplo, realizar mediciones (anemómetro, anemoclimómetro, anemógrafo, etc.).

A partir de esta base se puede elaborar una jerarquía en donde cada concepto se relacionará de manera semántica con otros y así de forma continuada. De esa manera cada categoría conceptual se encuentra representada por un conjunto de tipos de información que tiene en cuenta las características de un concepto y las entidades de esa categoría en el mundo real. Con estas categorías conceptuales (y sus respectivos esquemas) se modeló una base de conocimiento, estableciendo relaciones entre conceptos de "tipo de" (para relaciones de especialización), "parte de" (para las estructurales), "resultado de" (relaciones causa-efecto) o "atributo de" (asociativas). Así por ejemplo, el concepto "abanico fluvial" tiene una relación "tipo de", que es "abanico", una relación "atributo de" con "aluvial" y es "resultado de" del concepto "corriente".

Toda esta información semántica es almacenada en la propia base de conocimiento, a pesar de que la implementación actual de la misma haga un uso limitado de ella. No obstante, resulta especialmente relevante este hecho, principalmente por dos razones. En primer lugar debido a que es frecuente en la actualidad que numerosos proyectos almacenen su información en sistemas heredados (*legacy systems*) en los cuales cualquier cambio en su diseño podría resultar dramático. La posibilidad de completar y añadir información semántica en dichos sistemas permitiría acoplarlos (*mapping*) con formas de representación que sí hagan uso de dicha información, sin necesidad de realizar ninguna modificación en los sistemas actuales, y accediendo a su información de forma semántica. Por otra parte, también es reseñable el hecho de que el diseño de sis-

temas basados en formas de representación más ricas semánticamente, como las ontologías, es más complejo y por tanto requiere más tiempo para su elaboración.

"El diseño de sistemas basados en formas de representación más ricas semánticamente, como las ontologías, es más complejo y requiere más tiempo para su elaboración"

En la actualidad nos encontramos trabajando en la creación de una ontología que permita el acceso a la base de conocimiento utilizando la información semántica almacenada. No obstante, lo más interesante de esta propuesta radica en el hecho de que la información no se encuentre replicada en ambas representaciones. Para ello se están utilizando herramientas que permiten vincular ontologías a otras formas de representación, como *D2RQ* de la *Freie Universität Berlin*, consiguiendo una ontología que carga los datos necesarios en el momento de ser solicitados (por ejemplo por parte de un razonador), pero que no los almacena.

http://sites.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/

Por último, y una vez organizados todos los conceptos vinculándolos a varios idiomas, se elaboraron las entradas terminológicas de manera consensuada y supervisada por especialistas. La figura 4 muestra un ejemplo de entrada terminológica.

Como puede observarse en la entrada terminológica de ejemplo, ésta cuenta con un enlace a un banco de imágenes. Desde el comienzo del proyecto se pensó que, con fines didácticos y para lograr un mayor alcance en la explicación de los términos, sería interesante añadir contenido multimedia a la información asociada a cada término. Junto a la base de datos con la información referida al concepto (definición, equivalente en otros idiomas, relaciones con otros conceptos, categoría gramatical, contextos representativos, concordancias, etc.) se reservó un campo para albergar información multimedia. Dichos ficheros formarían un fondo de imágenes, vídeos, etc., que permitirían la creación de un tesauro visual de conceptos costeros. Además, esto posibilitaría crear una tipología de información gráfica que relacione ilustración y texto científico-técnico según criterios lingüístico-cognitivos.

Para poder gestionar toda esta gran cantidad de información y emplear un sistema de consulta eficaz fue necesario el uso de las diversas herramientas informáticas que se explican a continuación.

2.1. Desarrollo e implementación

Desde el punto de vista informático y de implementación del sistema de información, la experiencia previa fue fundamental a la hora de consolidar la arquitectura actual de *PuertoTerm*. Inicialmente el proyecto surgió como una memoria sobre términos especializados relacionados con el río Guadalfeo, información que era almacenada en una única base de datos centralizada. Entre sus funcionalidades destacaba un módulo constituido por varios formularios para su alimentación, y otro para la generación de informes que servía para la representación de la información contenida de un modo adecuado.

Sin embargo, a medida que la ambición del proyecto crecía, esta estructuración iba quedando obsoleta. El modelo centralizado obligaba a cada usuario a descargar la totalidad de la base de datos para realizar las modificaciones deseadas de forma local y posteriormente sobrescribir la versión disponible en el servidor.

Esto daba lugar a problemas de concurrencia. El sistema no facilitaba que varios usuarios pudiesen cooperar para realizar modificaciones de forma conjunta. Ni siquiera se garantizaba que cambios realizados de manera simultánea por diferentes usuarios se viesen reflejados finalmente en la base de datos. Eran los propios usuarios los encargados de comprobar que ningún otro usuario estuviera realizando modificaciones en un momento dado.

Otro aspecto fundamental que empezaba a verse afectado era el que concernía a la seguridad. Un usuario malintencionado que accediera a la base de datos podía eliminar o realizar cualquier modificación sin control alguno. Es más, era necesario realizar copias de seguridad de forma periódica para poder recuperar el sistema. La solución dada a este problema se basó en una configuración extraordinariamente restrictiva por parte del cortafuegos instalado en el servidor, que limitaba las posibilidades de acceso a los usuarios. Por ejemplo, sólo se podía acceder a la base de datos desde los ordenadores ubicados en los despachos de cada uno de los miembros del grupo, impidiéndoles trabajar en cualquier otro lugar.

Otro de los grandes problemas generados mediante esta forma de trabajo era que el único registro que se podía obtener del uso del servidor venía dado por el cortafuegos. Éste, sin embargo, no proporciona datos acerca del uso concreto realizado en el servidor, así que no se podían obtener estadísticas de uso u otro tipo de información que mejorase el aprovechamiento del servidor.

Por tanto, era necesario pasar de una arquitectura en donde el sistema de gestión de base de datos estaba orientado a un entorno personal, a un nuevo escenario

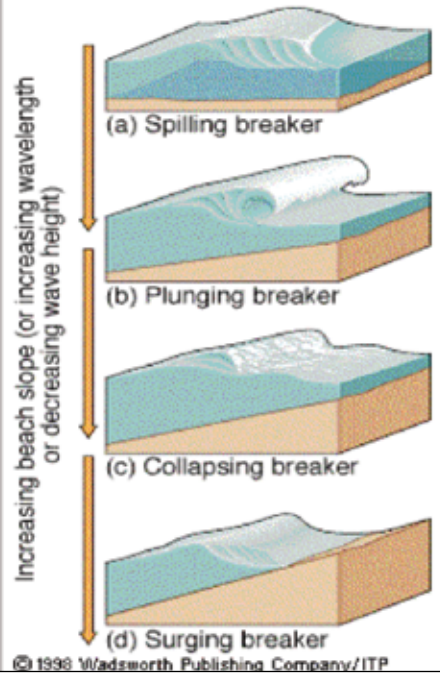
Rompiente de derrame	Ola que presenta un movimiento progresivo de atenuación y poca fuerza de ascenso en una onda que rompe dando lugar a la aparición de espuma, extendiéndose por una amplia zona de barrido o surf.	-Ola -Rompiente
Contexto	Entre las posibles razones de recuperación de la misma, puede postularse como factor importante la baja altura del oleaje cotidiano que caracteriza el tipo de rompiente de derrame ("spilling breaker") [vínculo a archivos multimedia], el cual tiene la característica de expandirse por una amplia zona de barrido o "surf", predominando la acreción hacia tierra por la caída libre de las partículas debido a que la ola tiene poca fuerza de ascenso, lo cual aunado al relativo abrigo topográfico podría explicar la recuperación de la playa. Fuente: *http://acta.ivic.ve/52-3/articulo5.pdf*	-Rompiente en voluta -Rompiente de colapso -Rompiente ondulada (Palabras en cursiva en azul indica hipervínculos)
Spilling breaker (inglés)	Breaker in which the bubbles and turbulent water spill down front face of a wave. The upper 25 percent of the front face may become vertical before breaking. Breaking generally occurs over quite a distance.	-Wave -Breaker
Contexto	*Breaker type* refers to the form of the wave at breaking. Wave breaking may be classified in four types: as spilling, plunging, collapsing, and surging In *spilling breakers* [vínculo a archivos multimedia], the wave crest becomes unstable and cascades down the shoreward face of the wave producing a foamy water surface. In *plunging breakers* [vínculo a archivos multimedia], the crest curls over the shoreward face of the wave and falls into the base of the wave, resulting in a high splash. In *collapsing breakers* [vínculo a archivos multimedia] the crest remains unbroken while the lower part of the shoreward face steepens and then falls, producing an irregular turbulent water surface. In *surging breakers* [vínculo a archivos multimedia], the crest remains unbroken and the front face of the wave advances up the beach with minor breaking. Fuente: *http://www.vulcanhammer.net/marine/Part-II-Chap_4entire.pdf*	-Plunging breaker -Collapsing Breaker -Surging breaker (Palabras en cursiva en azul indica hipervínculos)
Schwallbrecher (alemán)	Der Schwallbrecher tritt bei relativ großen Wellensteilheiten an flachen Stränden auf. Unter Beibehaltung der annähernd symmetrischen Form läuft das Wasser an der Oberfläche mit zunehmender Steilheit über den Wellenberg. Eine Brecherzunge löst sich aus dem Wellenberg, gleitet als Schwall den Leehang herunter und erzeugt zunehmend Wirbel und Turbulenzen.	-Welle -Brandungswelle
Contexto	Dabei wird die Wellenenergie über eine Strecke von mehreren Wellenlängen dissipiert.	-Sturzbrecher -Schwallbrecher -Reflexionsbrecher
	Beim Sturzbrecher, der für steilere Strände charakteristisch ist, wird der Wellenberg ›übersteilt‹ und kippt vornüber in das Wellental, wobei auch verfestigte Sedimente erodiert und grobe Kiese und Wacken bewegt werden. An flacheren Stränden überwiegen Schwallbrecher, bei denen der Wellenberg schäumend in das Wellental bricht. Letzlich wird in beiden Fällen Wellenenergie ›vernichtet‹ (Dissipation), indem sie auf Sedimentpartikel übertragen wird, im Falle des Sturzbrechers allerdings nur auf kurzer Strecke mit hoher morphogenetischer Wirksamkeit, beim Schwallbrecher auf einem breiten Streifen mit geringerer Wirksamkeit. Fuente: *http://www.uni-konstanz.de/llmnologie/wetlands/PDF-Dokumente/Ostendorp-2004-SVGB-Bd122.pdf*	(Palabras en cursiva en azul indica hipervínculos)
Banco de Imágenes	*archivo de audio spilling breaker* *archivo de video spilling breaker* archivo de imagen **SPILLING BREAKER:**	

Figura 4. Entrada terminológica del término rompiente de derrame

con varios usuarios actuando a la vez. En el caso de *PuertoTerm*, este punto resulta clave debido a la distribución espacial de sus integrantes. Para este cambio se aprovecharon las facilidades que ofrecen internet y sus numerosos marcos de trabajo disponibles.

Así, de un modelo centralizado y monolítico que hacía un uso ineficiente del servidor, se pasó a una arquitectura cliente-servidor, mucho más adecuada para un ambiente web. Un usuario ya no tendría que descargar la totalidad de dicha base de datos a su ordenador, aún cuando fuese a utilizar únicamente una parte, y provocar un escaso aprovechamiento del ancho de banda.

2.1.1. Portal de trabajo colaborativo

La arquitectura (figura 5) que se creó *ad hoc* está formada por diversos módulos:

a. Entrada de información: constituido por una serie de formularios web que se emplean para introducir la información relativa a los conceptos, términos, recursos y sus interrelaciones. Contiene varias tecnologías que hacen su utilización más amigable y eficaz, tales como atajos de teclado o mecanismos de búsqueda mientras se escribe, independiente del navegador web empleado.

b. Validación: revisa y corrige la información almacenada en la base de conocimiento. Incluye una funcionalidad de búsqueda que permite acceder de forma directa a cualquier elemento y que se puede realizar en base a cadenas completas o subcadenas, lo que amplia sus posibilidades. Debido al gran volumen de información introducido, no se requiere que genere una consulta sensible al caso. Por último, éste módulo facilita diferentes opciones para la presentación de la información. Así por ejemplo, se puede elegir si se valida una serie de registros mediante una navegación secuencial por orden de introducción, y otra ordenada mediante criterios alfabéticos.

c. Gestión de las comunicaciones: permite asociar discusiones a un concepto o término concreto. Otra posibilidad es la de generar informes de forma totalmente personalizada, para distribuir contenido selectivo a personas ajenas al sistema. De esta forma se facilita la interacción con usuarios a las que no se les desea dar acceso a la totalidad de la información. Finalmente, también se puede distribuir cualquier tipo de información mediante un simple gestor de contenidos integrado a cualquier visitante de la plataforma, que puede interactuar proporcionando comentarios acerca de la misma.

d. Usuarios: hace posible gestionar los accesos al sistema y la base de conocimiento. De esta manera se almacena cierta información asociada al usuario cada vez que accede y con el objeto de localizar posibles problemas. Cada usuario tiene restringida la información que puede manipular según su perfil.

e. Seguridad: proporciona mecanismos de recuperación en caso de aparición de problemas. Un agente actúa realizando copias de seguridad tanto interna como externamente en el servidor. Debido al volumen de información, tiene en cuenta la actividad de los usuarios para, por ejemplo no actuar si no han realizado ningún cambio. Al mismo tiempo, dos sistemas independientes se encargan de monitorizar de forma continua el correcto estado del servidor que contiene todo el sistema.

2.1.2. Base de datos *versus* ontología

El cambio de naturaleza de la información residente en *PuertoTerm*, pasando de meramente organizar la estructura e integridad de sus datos, a tratar de especificar el significado del dominio de la ingeniería de costas mediante conceptos y sus interrelaciones, también condicionó la elección del nuevo modelo. Esta nueva forma de representación se acerca mucho más a las bases de conocimiento y ontologías, que a las bases de datos. Las enormes posibilidades que abre la representación de información en este tipo de modelos, como la bien conocida web semántica, hacía esta elección muy atractiva.

Sin embargo, la mayoría de las herramientas disponibles para el desarrollo de ontologías están pensadas aún de forma monolítica. La idea subyacente es realizar pequeñas ontologías que puedan integrarse y consensuarse entre individuos. No obstante, la naturaleza de una ontología de las características de *PuertoTerm*, que necesita ser alimentada por muchos usuarios, no se ve reflejada en ese estereotipo. En la actualidad, diversos proyectos pretenden cubrir este hueco, como por ejemplo *Collaborative Protégé*, un añadido (*plugin*) para el conocido editor de ontologías *Protégé* que permite la edición colaborativa de las mismas. Sin embargo, como sus propios desarrolladores apuntan, en la actualidad se trata de un prototipo bajo desarrollo.

http://protege.stanford.edu/

http://smi-protege.stanford.edu/collab-protege/collab-protege.html

Otro de los puntos en contra para la elección de este tipo de herramientas fue la escasa usabilidad que ofrecen. Están orientadas hacia usuarios experimentados, y su proceso de aprendizaje consume gran cantidad de tiempo, que no podía ser asumida por parte del proyecto.

Por tanto, la única solución viable pasaba por implementar una interfaz propia que accediese de forma remota a la información residente en la ontología. Siendo una opción que, aunque cumpliendo todos los requerimientos mencionados, también consumiría una gran cantidad de tiempo.

Figura 5. Aspecto de la interfaz creada para la introducción de datos y comunicación entre los miembros del grupo

Por todas estas razones, se decidió mantener un sistema de gestión de base de datos para almacenar la base de conocimiento. La amplia utilización de las bases de datos ha permitido una gran optimización en su rendimiento, algo que aún no ha sucedido con las herramientas ontológicas, que se encuentran en un grado incipiente.

La información semántica que no puede ser empleada por la utilización de un modelo más restrictivo en lo que respecta al nivel de representación que es capaz de expresar (es decir, la base de datos), es también almacenada con el objeto de utilizarla para establecer una correspondencia posterior con modelos más ricos en su semántica (en nuestro caso, una ontología). De esta manera, se alimenta un sistema por medio de un mecanismo ampliamente probado, sin que eso menoscabe su capacidad de representación ni implique duplicación de la información. La figura 6 muestra las relaciones existentes dentro de la base de datos.

2.1.3. Recuperación y visualización de la información

Desde el principio se pensó que el acceso a la información almacenada en la base de datos se debería realizar por medio de consultas (*querying*) y búsqueda analítica (*browsing*). Teniendo en cuenta que toda la información se encuentra centralizada en varias tablas almacenadas en *MySQL*, el acceso a la misma no debería ser excesivamente complejo y, además, posibilitaría fáciles migraciones con vistas a integrar este trabajo en otros proyectos de parecidas características. Además, se planteó la posibilidad de mostrar la información de una manera más visual y menos lineal. Por ese motivo, se optó por la utilización de tecnologías de visualización de la información en 2 y/o 3 dimensiones que faciliten la rápida comprensión de las estructuras arbóreas inherentes dentro de cualquier sistema clasificatorio.

"La utilización de tecnologías visuales para mostrar la información recuperada y las relaciones existentes resulta de especial trascendencia en *PuertoTerm*"

La utilización de tecnologías visuales para mostrar la información recuperada y las relaciones existentes resulta de especial trascendencia en *PuertoTerm* ya que posibilita una mejor y más profunda compresión del contenido almacenado, proporcionando mecanismos para acceder a información relevante que podría no ser accesible mediante otros sistemas de representación. Además de permitir inferir conocimiento de manera más sencilla que en otros sistemas tradicionales.

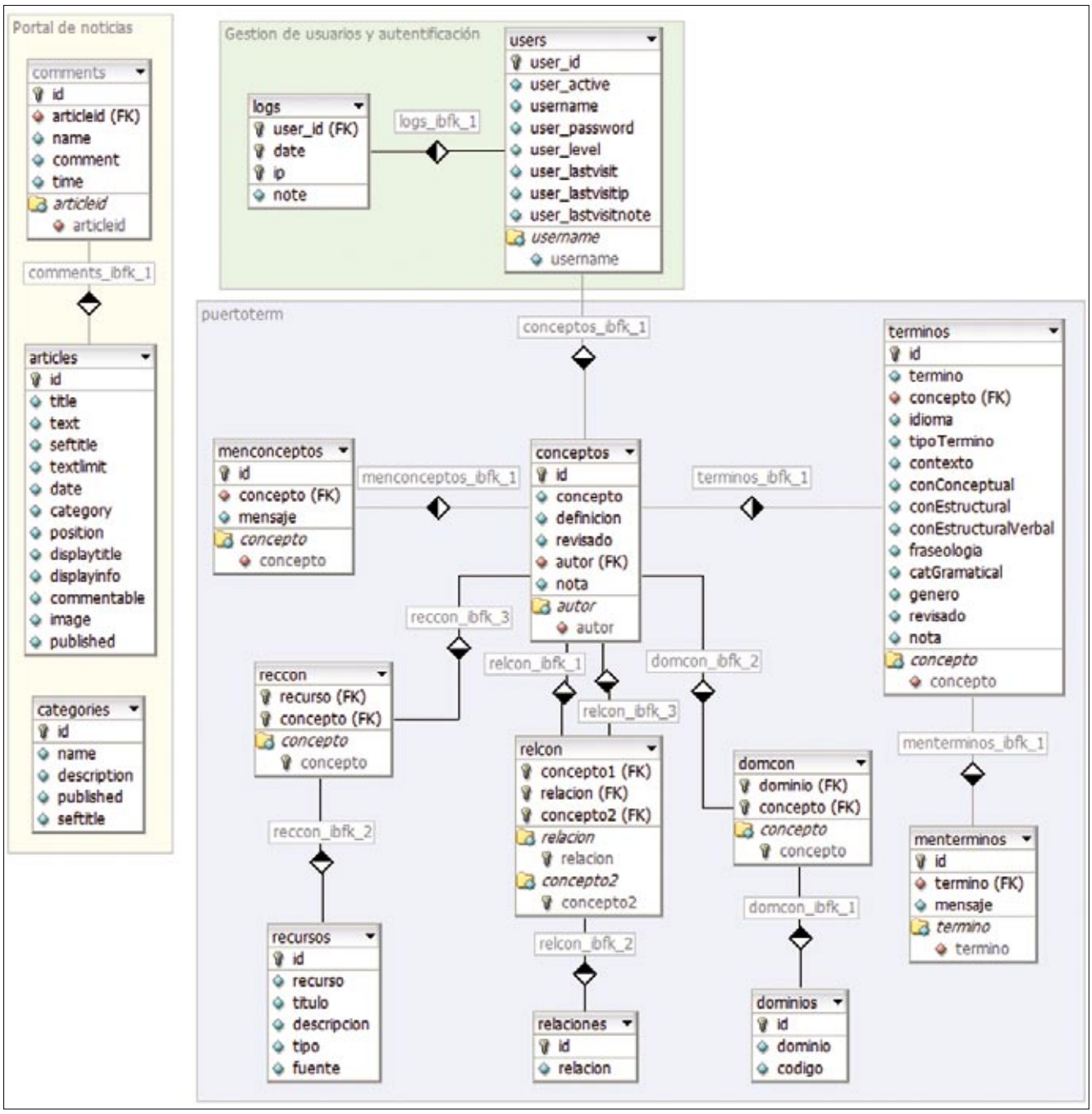


Figura 6. Relaciones existentes dentro de la arquitectura de PuertoTerm

No obstante, la presentación de información a través de la web cuenta con numerosos problemas, debido a las limitaciones del lenguaje en el que está implementada. Html limita el acceso a la información a modelos bidimensionales con escasa información estructurada. De ahí surge la necesidad de introducir algún complemento para que los navegadores extiendan su funcionalidad visual.

De entre todas las opciones disponibles la más difundida y con más posibilidades son los *applets*, ya que están implementados en un lenguaje muy potente como es *Java*. Las opciones de un lenguaje de programación tan versátil son prácticamente ilimitadas, presentando características para interactuar con otros sistemas, la optimización de cualquiera de sus componentes, o el desarrollo de cualquier funcionalidad a medida.

Además, su amplia portabilidad es una característica fundamental para el desarrollo de cualquier sistema que quiera ser usable. El amplio rango de usuarios a los que *PuertoTerm* pretende abarcar hace de ésta una de sus principales características. Los usuarios están divididos fundamentalmente en tres grupos. El primero corresponde a los lingüistas que alimentan y desarrollan la base de conocimiento. El grupo de expertos en ingeniería de costas conforman el segundo de ellos. Finalmente, los usuarios finales del sistema podrían abarcar desde la administración pública hasta particulares que

deseen ampliar sus conocimientos mediante la lectura de textos en lenguas foráneas.

Las técnicas de visualización existentes para bases de datos no tienen en cuenta la información semántica considerada en esta aplicación, por lo que no son válidas para nuestro propósito. Por otra parte, las técnicas empleadas dentro de ese campo para las ontologías son más ricas en expresividad, pero aún son muy limitadas, y generalmente están integradas en el propio editor donde se desarrolla la ontología (*Jambalaya*, *Knowledge Tree*, *Ontoviz*, *TGViz*, etc.).

Por tanto, era necesario utilizar alguna técnica propia para la visualización del conocimiento. Una primera consideración podría llevar a pensar en el uso de *Mindmaps*, *Concept Maps*, y otra serie de representaciones de la información. Sin embargo, cada una de ellas está pensada para un tipo de aplicaciones concretas. Así, los *Mindmaps* están fundamentalmente orientados a caracterizar el conocimiento que un individuo tiene acerca de un tema principal, estructurando dicho contenido en forma de árbol y jerarquías radiales, no teniendo las conexiones entre sus elementos ningún tipo de restricción asociada. Los *Concept Maps* están enfocados a visualizar las relaciones entre conceptos (**Novak**, 1991). No tiene por qué haber un elemento principal en ellos, y la información acerca de sus conexiones es mucho más expresiva que en los *Mindmaps*, pero seguía sin permitirnos mostrar todo el potencial de las relaciones de *PuertoTerm*.

De esta forma se optó por un solución intermedia, el software *ThinkMap*, que combina lo mejor de ambas representaciones de una manera eficaz para el desarrollo de nuestra aplicación (**Kroeker**, 2004). Es empleado con notable éxito en productos comerciales de similares características a *PuertoTerm*, como *Visual Thesaurus*, *Sony Music Licensing* o la *National Oceanic and Atmospheric Administration* de los EUA. Ayuda a organizar la información dotándola de sentido, ya que se presenta de forma contextual, tomando un elemento como central, y proporcionando toda la información necesaria para caracterizar sus relaciones.

http://www.thinkmap.com

http://www.visualthesaurus.com

http://www.sonymusic.com/licensing

http://www.noaa.gov/

El programa facilita una infraestructura (*framework*) que permite desarrollar aplicaciones muy flexibles (con un innovador entorno) extremadamente eficientes, fácilmente escalables y que puedan interactuar con diferentes fuentes de información. Éste último hecho adquiere especial relevancia en *PuertoTerm*, ya que aunque en estos momentos los *ThinkMaps* implementados obtengan toda la información a partir de la base de datos (incluida la semántica), en nuestros futuros trabajos está la idea de que se alimente directamente de la ontología, algo que sería mucho más natural, siendo una opción no disponible en otros sistemas.

ThinkMap emplea un conjunto de algoritmos que permiten mostrar grandes cantidades de información de manera eficiente por medio de una interfaz bidimensional. De esa forma es posible comprender fácilmente la relación existente entre cientos de nodos de información. Además, la representación mostrada no es estática, ya que el usuario puede en todo momento realizar continuas búsquedas a partir de los resultados de una primera. Conforme se van realizando, los gráficos se regeneran, creciendo o decreciendo la cantidad de nodos mostrados en función del número de relaciones que existan entre los nuevos términos localizados. La figura 7 muestra las diversas opciones de búsqueda, historial, y representación de la información que se han implementado en *PuertoTerm* empleando esta herramienta.

Otro aspecto de especial relevancia para nuestros propósitos radica en la facilidad que tiene el programa para integrar contenidos multimedia. Este nuevo tipo de información visual complementa y amplia el contenido textual, como queda atestiguado en numerosos ejemplos (*Le grand dictionnaire terminologique*) o la inclusión de opciones para la gestión de contenido multimedia en la mayoría de las herramientas para la gestión terminológica.

http://www.granddictionnaire.com

3. Conclusiones

La gestión eficiente del conocimiento dentro de un dominio a partir de la utilización de herramientas y metodologías adecuadas resulta requisito imprescindible en una sociedad en donde la información juega un lugar tan destacado. Sin embargo, para que el conocimiento sea efectivo debe no sólo poder estructurarse para un fin particular, sino que debe poder combinarse e integrarse con otros sistemas de similares propósitos, para que dicho conocimiento sea enriquecido.

Éste es uno de los principales objetivos de la denominada web semántica, poder proporcionar una infraestructura común para la representación del conocimiento y, de esta manera, que la información no sea propietaria de las aplicaciones que la utilizan. En la actualidad la web está orientada al intercambio de documentos entre humanos. Al utilizar para la gestión del conocimiento tecnologías más expresivas en su semántica, se facilitaría además el intercambio y la inferencia de nuevo conocimiento por parte de las propias máquinas.

Figura 7. Aspecto de la interfaz de recuperación de la información de PuertoTerm. La pantalla principal (A) la divide en dos partes. En la izquierda se encuentran las funciones para la búsqueda de la información, el mapa que muestra las relaciones entre los términos y las definiciones (B), el histórico de búsquedas (C) y la lista de resultados. El mapa que se genera de forma automática se relanzará con una nueva búsqueda cada vez que seleccionemos alguno de los resultados del mapa original (E). En la derecha se reúne la información relativa al dominio al que pertenece el término de búsqueda (F) que, por supuesto, también se puede emplear para búsquedas por browsing. Toda esta información se completa con el equivalente idiomático del término en diversas lenguas y de un conjunto de recursos gráficos (G) que facilitan la comprensión de la idea expuesta

"La metodología seguida en este trabajo se puede extender a otras áreas en donde las aplicaciones que se desean desarrollar compartan los objetivos, adaptando las necesidades específicas"

Desde el primer momento el proyecto fue consciente de esta nueva forma de desarrollar aplicaciones para la gestión del conocimiento específico. A pesar de no utilizar tecnologías propias de la web semántica desde el principio, el diseño inicial tenía en cuenta la necesidad de integrar en ella todo el contenido de *PuertoTerm*, por lo que en la actualidad su información es accesible mediante el estándar propuesto *OWL* (*on-*

tology web language) (**Smith**, 2004). De esta manera el conocimiento de *PuertoTerm* se podría relacionar con aplicaciones llevadas a cabo dentro de su área.

Aunque la ingeniería de costas y puertos es una disciplina con ciertas peculiaridades (como ya hemos comentado) la metodología seguida en este trabajo se puede extender a otras áreas en donde las aplicaciones que se desean desarrollar compartan los objetivos, adaptando las necesidades específicas. La reutilización y combinación de muchas de las herramientas disponibles en la actualidad para la creación de aplicaciones ha propiciado la construcción de una infraestructura completa sin que ello haya supuesto hacerlo desde cero, algo que es fácilmente extrapolable.

En la actualidad nos encontramos completando y mejorando la calidad de la información contenida en *PuertoTerm*. Esta tarea se verá rematada por el diseño de una herramienta visual de consulta semántica que aproveche las características de *PuertoTerm* directamente desde la ontología y no desde la semántica almacenada en la base de datos. Como hemos comentado anteriormente, este hecho no supondría cambiar de herramienta de visualización, con el consiguiente ahorro de aprendizaje asociado para los usuarios. Por último también se plantea la posibilidad de incorporar nuevas tecnologías emergentes en la web semántica, tales como los microformatos, *RDFa* y *Grddl* (*gleaning resource descriptions from dialects of languages*). Todas estas herramientas están orientadas a la inclusión de la semántica en las propias páginas web, permitiendo que la base de conocimiento no permanezca oculta en un segundo plano.

4. Nota

1. Esta investigación forma parte del proyecto "Marcos de conocimiento multilingües en la gestión integrada de zonas costeras" *[MarcoCosta](P06-HUM-01489)*, subvencionado por la *Junta de Andalucía*.

5. Bibliografía

Berners-Lee, T.; **Hendler, J.**; **Lassila, O.** "The semantic web". En: *Scientific American magazine*, 2001, v. 284, n. 5, pp. 34-43.

Faber, P., et al. "Process-oriented terminology management in the domain of coastal engineering". En: *Terminology*, 2006, v. 12, n. 2, pp. 189-213.

Faber, P.; **Mairal-Usón, R.** *Constructing a lexicon of English verbs*. Berlin: Mouton de Gruyter, 1999.

Faber, P.; **Márquez, C.**; **Vega, M.** "Framing terminology: a process-oriented approach". En: *Meta*, 2005, v. 50, n. 4. Consultado en: 12-03-07. *http://www.erudit.org/livre/meta/2005/000255co.pdf*

Fillmore, C. "Frame semantics". En: *The Linguistic Society of Korea* (ed.). *Linguistics in the morning calm.* Seoul: Hanshin, 1982, pp. 111-137.

Fillmore, C. F.; **Johnson C. R.**; **Petruck, M. R. L.** "Background to FrameNet". En: *Journal of lexicography*, 2003, v. 16, n. 3, pp. 235-250.

Fillmore, C.; **Atkins, S.** "FrameNet and lexicographic relevance". En: *Proceedings of the Granada conference on linguistics*.

Gahl, S. "Automatic extraction of subcategorization frames for corpus-based dictionary making". En: *Euralex'98 proceedings*, 1998, pp. 445-452.

Irazazábal, Amelia. "Terminología y documentación". En: *Jornada panllatina de terminologia: perspectives i camps d'aplicació.* 1996. Consultado en: 08-03-07.
http://www.realiter.net/jorb/irazazabal.htm

Kroeker, Kirk L. "Seeing data: new methods for understanding information". En: *IEEE computer graphics and applications*, 2004, May/Jun, v. 24, n. 3, pp. 6-12.

López-Rodríguez, Clara-Inés; **Tercedor-Sánchez, Maribel**; **Faber-Benitez, Pamela**. "Gestión terminológica basada en el conocimiento y generación de recursos de información sobre el cáncer: el proyecto Oncoterm". En: *RevistaeSalud.com*, v. 2, n. 8, 2006. Consultado en: 07-03-07.
http://www.revistaesalud.com/index.php/revistaesalud/article/view/127/322

Lyons, J. *Semantics*. London: Cambridge University Press, 1977.

Martín-Mingorance, L. "Functional grammar and lexematics". En: **Tomaszczyk, J.**; **Lewandowska, B.** (eds.). *Meaning and lexicography*. Amsterdam/Philadelphia: John Benjamins, 1989, pp. 227-253.

Martín-Mingorance, L. "Lexical logic and structural semantics: methodological underpinnings in the structuring of a lexical database for natural language processing. En: *Hoinkes* (eds.). *Panorama der Lexikalischen Semantik*. Tubinga: Gunter Narr, 1995, pp. 461-474.

McEnery, A.; **Wilson, A.** *Corpus linguistics*. Edimburgo: Edinburgh University Press, 1996.

Novak, J. D. "Concept maps and Vee diagrams: two metacognitive tools to facilitate meaningful learning". En: *Instructional science*, 1991, n. 19, pp. 1-25.

Smith, M. K.; **Welty, C.**; **McGuinness, Deborah L.** *OWL web ontology language guide*. *W3C Recommendation*, 2004.
http://www.w3.org/TR/owl-guide/

Pérez-Álvarez-Ossorio, J. R. *Introducción a la información y documentación científica*. Madrid: Alhambra/Universidad, 1998.

Pinto, María; **Cordón, José-María** (eds.). *Técnicas documentales aplicadas a la traducción*. Madrid: Síntesis, 1999.

Puertos del Estado. *Recomendaciones para obras marítimas. ROM. 0.0 Procedimiento general y bases de cálculo en el proyecto de obras marítimas y portuarias*. Madrid: Ministerio de Obras Públicas y Urbanismo, 2001, 220 págs. ISBN 84-88975-30-9.

Sales-Salvador, Dora. *Documentación aplicada a la traducción: presente y futuro de una disciplina*. Gijón: Trea, 2006.

ISO-5964. *Guidelines for the establishment and development of multilingual thesauri*. Geneva: International Organization for Standardization, 1985.

***Jose A. Senso**, Departamento de Biblioteconomía y Documentación, Universidad de Granada.*
jsenso@ugr.es

***Pedro-Javier Magaña-Redondo**, Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada.*
pedro@correo.ugr.es

***Pamela Faber-Benitez**, Departamento de Traducción e Interpretación, Universidad de Granada.*
pfaber@ugr.es

***María-Amparo Vila-Miranda**, Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada.*
vila@decsai.ugr.es

Características y difusión de las revistas científico-técnicas españolas de ciencias de la actividad física y el deporte

Por **Miguel Villamón-Herrera, José Devís-Devís, Alexandra Valencia-Peris** y **Javier Valenciano-Valcárcel**

Resumen: *El objetivo de este artículo es presentar la situación actual de las revistas españolas de Ciencias de la Actividad Física y el Deporte (CCAFD). Para ello se actualiza el inventario de publicaciones periódicas vigentes y se analizan varias de sus características editoriales básicas como la antigüedad, periodicidad, lugar de edición, institución editora, tipo de soporte y tipo de formato editorial. También se estudia la circulación de dichas publicaciones en las bases de datos nacionales e internacionales selectivas y la difusión que tienen por internet a través de la accesibilidad, el contenido y los servicios añadidos que ofrecen. Se concluye que las CCAFD es un campo joven y poco consolidado académicamente, según indican las características editoriales básicas de sus revistas científico-técnicas. Sin embargo, dichas publicaciones han experimentado una importante mejora en los últimos años en su indización en bases de datos y en su presencia en internet.*

Palabras clave: *Ciencias de la actividad física y el deporte; Revistas españolas; Características editoriales; Bases de datos; Circulación; Visibilidad; Internet.*

Miguel Villamón *es doctor en Educación Física y profesor del Departamento de Educación Física y Deportiva de la Universitat de València, donde imparte docencia desde 1991. Sus líneas de investigación se centran en la evaluación de las publicaciones científicas y la enseñanza de los deportes, temas sobre los que es autor de diversos artículos y monografías.*

José Devís *es doctor en Ciencias de la Educación y profesor Titular de Universidad en la Facultad de Ciencias de la Actividad Física y el Deporte de la Universitat de València. Sus líneas de investigación son la evaluación de la ciencia, y la pedagogía de la actividad física y la salud. Ha coordinado el libro 'La educación física, el deporte y la salud en el siglo XXI' y es autor de, 'Educación física, deporte y currículum'. Forma parte del consejo editorial de las revistas Sport, Education and society, Physical education and sport pedagogy, y Movimento.*

Alexandra Valencia *es estudiante de Ciencias de la Actividad Física y el Deporte y Magisterio, becaria-colaboradora de la Unidad de Investigación de Teoría y Pedagogía de la Actividad Física y el Deporte de la Universitat de València.*

Javier Valenciano *es licenciado en Educación Física y Primer Premio nacional fin de carrera. Ha sido becario FPI del Departamento de Educación Física y Deportiva de la Universitat de València y se ha dedicado al estudio de la calidad y de la producción científica de las revistas españolas de ciencias de la actividad física y el deporte, temas sobre los que ha publicado varios artículos.*

Title: Dissemination and characteristics of Spanish physical activity and sport sciences scientific and technical journals

Abstract: *The aim is to present the current situation of Physical Activity and Sport Sciences (PASS) journals in Spain. The inventory of Spanish journals is updated and several basic editorial characteristics, such as longevity, periodicity, place of publication, editorial institution and support, and type of editorial format are analyzed. The journals' circulation in national and international databases and internet dissemination (based on accessibility, content and added services) are also studied. We conclude that PASS is still a young field that has not yet consolidated its academic presence, as shown by the basic editorial characteristics of it scientific and technical journals. Nonetheless, these journals have seen a significant improvement in their database indexing and internet presence in recent years.*

Keywords: *Physical activity and sport sciences; Spanish journals; Editorial characteristics; Database; Circulation; Visibility; Internet.*

Villamón-Herrera, Miguel; **Devís-Devís, José**; **Valencia-Peris, Alexandra**; **Valenciano-Valcárcel, Javier**. "Características y difusión de las revistas científico-técnicas españolas de ciencias de la actividad física y el deporte". En: *El profesional de la información,* 2007, noviembre-diciembre, v. 16, n. 6, pp. 605-615.

DOI: 10.3145/epi.2007.nov.07

Introducción

Las revistas científicas son fundamentales en el ciclo de la ciencia por el papel que desempeñan en la divulgación de los resultados de investigación y porque son el canal de comunicación más utilizado y reconocido por la comunidad científica. Además resultan primordiales para el desarrollo científico de cualquier campo de conocimiento, siendo su número y calidad indicadores de la madurez del mismo. También sirven para dinamizar a la comunidad académica y se convi-

Artículo recibido el 01-10-07
Aceptación definitiva: 08-10-07

erten en elementos de referencia para la evaluación de la actividad investigadora de sus miembros.

En el caso de las Ciencias de la Actividad Física y el Deporte (CCAFD) se observa una mayor presencia de publicaciones periódicas en los circuitos internacionales. Desde hace varias décadas las revistas son indizadas por prestigiosas bases de datos (bdds) internacionales, y un grupo considerable de ellas se recogen en el *Journal citation reports* (ver **Hopkins**, 2004, 2005, 2007). Incluso han ido emergiendo bdds específicas como son *Spolit*, *Sport discus* y *Heracles* (**Aquesolo**, 2000). Sin embargo tienen aún poca visibilidad las revistas españolas y sus trabajos, a pesar del aumento de prestigio académico del mencionado campo en nuestro país. Al menos, no es comparable al aumento de productividad científica y a los recursos que las agencias nacionales y autonómicas han concedido a los investigadores desde la década de 1990 en que se consolidó el estatus académico y universitario de las CCAFD.

"A pesar del aumento de la productividad científica en las CCAFD, las revistas españolas del campo tienen poca visibilidad"

La falta de estudios sobre las publicaciones periódicas de las CCAFD, a excepción de unos primeros trabajos (**Devís**, **Antolín**, **Villamón**, **Moreno** y **Valenciano**, 2003; **Devís**, **Villamón**, **Antolín**, **Valenciano** y **Moreno**, 2004; **Villamón**, **Devís** y **Valenciano**, 2005; **Villamón**, **Devís** y **Valenciano**, 2006), no permite encarar las estrategias de mejora necesarias que ayuden a cambiar la situación anterior. Además, esos trabajos necesitan ser actualizados y completados para cumplir satisfactoriamente con dicha tarea. Por ello resulta fundamental abordar el análisis de las revistas científico-técnicas de CCAFD y, de esta manera, conocer en profundidad la situación actual de sus principales órganos de comunicación científica. Concretamente, en este trabajo se analizan las características editoriales básicas y la difusión, en particular la presencia en bdds y el contenido ofrecido en internet de las revistas vigentes en España. La rápida evolución de las publicaciones académicas, debido fundamentalmente al avance tecnológico, y la gran variedad de soportes documentales que existen actualmente (**Osca**; **Mateo**, 2003) así lo aconsejan. Pensemos, por ejemplo, en la repercusión editorial que implica la inmediatez electrónica y la posibilidad de acceso universal a los contenidos científicos que representa la iniciativa *open access*, sin restricciones económicas y de tiempo. Además, en última instancia la información proporcionada por este estudio resulta muy necesaria para todos los agentes participantes, directa o indirectamente, en los distintos momentos del proceso de edición de publicaciones periódicas. Es decir, desde los autores que podrán seleccionar informadamente la revista que mejor se adecue a sus necesidades, hasta los editores que pueden obtener información relevante para mejorar sus productos editoriales o los responsables de la política científica encargada de ayudar al mantenimiento y la internacionalización de las revistas científicas españolas.

Métodos

Para conocer la situación de las revistas españolas de CCAFD se requiere disponer de un listado de las que están vigentes. Por este motivo, el primer requisito metodológico consiste en elaborar un inventario reciente. Si bien **Devís** et al. (2003) hicieron un primer inventario con 26 revistas, la constante dinámica de creación y cese de revistas, así como la conveniencia de presentar la situación más reciente del campo de estudio, aconsejaban ofrecer un nuevo listado actualizado de 2007. Esta tarea se realizó a través de la consulta de las siguientes fuentes de información: bases de datos *ISSN*, *ISOC* e *IME*; catálogos colectivos *Rebiun*, *Catálogo colectivo de publicaciones periódicas* de la *Biblioteca Nacional de España* y *C17* del *Instituto de Salud Carlos III*; directorio *Latindex*; servicio de publicaciones *Dialnet*; buscadores *Google* y *AltaVista* y metabuscador *MetaCrawler* (que incluye los buscadores *Google*, *Yahoo!*, *MSN*, *Ask Jeeves*, *About*, *MIVA* y *LookSmart*). También se emplearon los mismos criterios de selección de actualizaciones anteriores (disponer de ISSN, estar activa al menos un año, tener carácter científico-técnico y ajustarse a la temática del campo).

"La existencia de un elevado número de revistas (32 en 2007) disminuye la competencia entre los originales que aspiran a publicarse, lo que puede abrir las puertas a trabajos de menor calidad"

Una vez conocidas las revistas objeto de estudio, se recogieron los datos relativos a las características editoriales y la difusión de las revistas vigentes. En el caso de las editadas en papel se consultaron directamente los últimos fascículos disponibles en las bibliotecas de la *Universitat de València* a finales de junio de 2007. Para las revistas electrónicas se analizaron los ejemplares digitales presentes en su web entre los meses de abril y junio del mismo año. En los casos en que fue necesario,

estos procedimientos se ampliaron a la consulta de las bases de datos bibliográficas mencionadas, a internet y a expertos.

De cada una de las publicaciones se comprobó una serie de datos (título, ISSN, fecha de inicio, lugar de edición, editorial, periodicidad, tipo de soporte, circulación en bases de datos y grado de difusión en internet), de acuerdo con otros trabajos recientes sobre este tema (**Abad**, **González**, **Martínez**, 2005; **Gómez del Pulgar**, 2006).

Resultados y discusión

Este apartado está organizado conforme a los dos objetivos del trabajo, las características editoriales básicas y la difusión de las revistas científico-técnicas españolas de CCAFD.

1. Características editoriales básicas

La actualización del inventario de publicaciones periódicas científico-técnicas españolas de CCAFD ha proporcionado un listado de 32 revistas vigentes en 2007, que se muestran en la tabla 1. Se trata de un número similar al de otros campos en España como, por ejemplo, antropología que posee 29 publicaciones, urbanismo con 35 o las 33 revistas de ciencias agrarias (**Román**, **Vázquez**, **Urdín**, 2002; **Urdín**, **Vázquez**, **Román**, 2003). Sin embargo, esta aparente buena salud del campo de estudio no se corresponde con la tradición universitaria e investigadora que poseen los campos anteriores y, probablemente, tampoco con la cantidad de miembros de la comunidad científica. El mantenimiento de este elevado número de revistas puede acabar afectando negativamente a la calidad de las mismas porque requiere disponer de muchos artículos originales. Además, la competencia entre dichos artículos para llegar a ser publicados es considerablemente menor de lo que exigiría un número inferior de revistas.

"El 56,2% de las revistas se editan en formato papel y electrónicamente, mientras que el 28,1% lo hace sólo en soporte papel y el 15,6% sólo electrónicamente"

El tipo de soporte en que se editan estas publicaciones periódicas muestra que sólo en papel existen 9 revistas, otras 18 lo hacen en soporte electrónico y en papel y 5 se editan únicamente en formato electrónico. Esta distribución indica una importante diversificación del soporte de edición que hace tan sólo una década era exclusivamente en papel. Si se compara esta distribución con la existente en el primer estudio de las revistas del campo correspondientes al año 2000 (**Devís** et al., 2003) se advierte que, del 7,7% de publicaciones editadas en versión exclusivamente electrónica, se ha pasado al 15,6%. Pero donde se observa un cambio muy significativo es en la categoría de 'papel y electrónico', pues en 2000 no había ninguna publicación que simultaneara ambos formatos y en 2007 ascienden hasta al 56,2%. Asimismo, destaca la transformación de 2 revistas a edición *online* gratuita que antes de 2007 era de pago y editadas en soporte papel: *Apunts. Medicina del deporte* y *Retos*. En definitiva, por lo que respecta al tipo de soporte, se observa un crecimiento considerable de la edición electrónica, aunque no sea en exclusividad, que se acompaña en muchos casos de la gratuidad de sus artículos o contenidos en *pdf*, como ocurre en otros campos en España e internacionalmente (**Vázquez**, **Urdín**, **Román**, 2003; **Aliaga**, **Suárez**, 2002).

Por otra parte se observa un crecimiento de 6 títulos respecto al inventario del año 2000. Han desaparecido 11 por cese, involución de sus contenidos científicos o retraso en su publicación, mientras que otros 12 han pasado a formar parte del actual inventario. Es, por tanto, el reflejo del dinamismo editorial que durante los últimos años parece mostrar un especial interés por las publicaciones periódicas. Así lo indica también el hecho de que cerca del 50% se hayan creado en los últimos 10 años (tabla 1). Sólo 11 publicaciones (34,4%) son anteriores a 1991, año en que se iniciaron los cursos de doctorado específicos de CCAFD y se intensificó la actividad investigadora, cuestiones éstas íntimamente ligadas a la creación y desarrollo de revistas científicas. No obstante, el número de altas y bajas en las publicaciones también indica una clara dificultad por su continuidad, lo que exige pensar en estrategias de consolidación o, como también se ha sugerido, de integración de títulos (**Aleixandre-Benavent**, **Valderrama-Zurián**, **González-Alcaide**, 2007; **Giménez**, **Gómez**, **Vázquez**, 2001; **Krauskopf**, **Vera**, 1995).

"La antigüedad media de las publicaciones periódicas es de 12,4 años y cerca de la mitad se ha creado en los últimos 10 años"

En un análisis detallado de la temática se advierte que predominan las revistas multidisciplinares, que suman 18 (56,2%) y poseen un contenido temático diverso relativo a la educación física (4), a las ciencias de la actividad física y el deporte (9), natación y actividades

Título	ISSN	Inicio	Lugar	Editor	Periodicidad
Agua y gestión	1131-8775	1987	Esplugues de Llobregat (Barcelona)	SEAE	Trimestral
Aloma. Revista de psicologia, ciències de l'educació i de l'esport	1138-3194	1997	Barcelona	Blanquerna, Universitat Ramon Llull	Semestral
Apunts. Educación física y deportes	1577-4015	1985 (*)	Barcelona	INEFC/Generalitat de Catalunya	Trimestral
Apunts. Medicina de l'esport	0213-3717	1985 (*)	Barcelona	Societat Catalana de Medicina de l'Esport / Consell Català de l'Esport / Generalitat de Catalunya	Cuatrimestral
Archivos de medicina del deporte	0212-8799	1984	Pamplona	Federación Española de Medicina del Deporte (Femede)	Bimestral
Avances en traumatología, cirugía ortopédica, rehabilitación, medicina preventiva y deportiva	0214-4077	1971	Barcelona	Asepeyo/Puntex	Trimestral
Biomecánica	1135-2205	1992	Barcelona	Sociedad Ibérica de Biomecánica y Biomateriales	Semestral
Cairon. Revista de ciencias de la danza	1135-9137	1995	Alcalá de Henares (Madrid)	SP Universidad de Alcalá	No se indica (Anual)
Comunicaciones técnicas. Publicación de la ENE de la RFE de Natación	1135-111X	1988	Madrid	Federación Española de Natación, E.N. de E.	Trimestral
Cuadernos de psicología del deporte	1578-8423	2001	Murcia	F. de Psicología, Universidad de Murcia / Dirección Gral. de Deportes de la CARM	Semestral
Cultura, ciencia y deporte	1696-5043	2004	Murcia	Universidad Católica San Antonio	Semestral
Derecho deportivo en línea	1579-2668	2001	Barcelona	Rafael Alonso Martínez	No se indica (Anual)
Fútbol. Cuadernos técnicos	1135-2817	1995	Sevilla	Wanceulen E. Deportiva	No se indica (Cuatrimestral)
Habilidad motriz. Revista de las ciencias de la actividad física y el deporte	1132-2462	1992	Córdoba	Colegio de Licenciados en EF y en CCAFD de Andalucía	Semestral
Kronos: la revista científica de la actividad física y el deporte	1579-5225	2002	Madrid	Universidad Europea de Madrid	Semestral
MD. Revista científica en medicina del deporte	1698-9775	2005	Sevilla	Junta de Andalucía, Consejería de Turismo, Comercio y Deporte, Centro Andaluz de Medicina del Deporte	No se indica (Cuatrimestral)

Tabla 1. Características editoriales básicas de las revistas científico-técnicas españolas de CCAFD

Título	ISSN	Inicio	Lugar	Editor	Periodicidad
Motricidad. European journal of human movement	0214-0071	1995	Cáceres	Asociación Española de Ciencias del Deporte	Semestral
NSW. Natación, saltos / sincro y waterpolo	1136-0003	1979	Cartagena (Murcia)	Asociación Española de Técnicos de Natación	Trimestral
RendimientoDeportivo.com	1578-7354	2002	Ribaseca (León)	Juan Carlos Morante Rábago	Cuatrimestral
RED. Revista de entrenamiento deportivo	1133-0619	1987	La Coruña	Boidecanto	Trimestral
Retos. Nuevas tendencias en educación física, deporte y recreación	1579-1726	2002	Almería	Federación Española de Asociaciones de Docentes de EF	Semestral
Revista andaluza de derecho del deporte	1886-6220	2006 (**)	Sevilla	Junta de Andalucía, Consejería de Turismo, Comercio y Deporte, Secretaría General para el Deporte	Semestral
Revista de educación física	1133-0546	1985	La Coruña	Boidecanto	Trimestral
Revista de psicología del deporte	1132-239X	1992	Palma de Mallorca	Universitat de les Illes Balears y Universitat Autònoma de Barcelona	Semestral
Revista española de educación física y deportes	1133-6366	1949 Nueva etapa: 2004	Madrid	CGICO de Licenciados en EF y en CCAFD	Semestral
Revista Iberoamericana de psicología del ejercicio y del deporte. Riped.	1886-8576	2006	Sevilla	Wanceulen E. Deportiva	Semestral
Revista iberoamericana de psicomotricidad y técnicas corporales	1577-0788	2000	Las Rozas (Madrid)	Asociación de Psicomotricistas del Estado Español y Red Fortaleza de Psicomotricidad.	Trimestral
Revista internacional de ciencias del deporte. Ricyde.	1885-3137	2005	Madrid	Ramón Cantó Alcaraz	Trimestral
Revista internacional de medicina y ciencias de la actividad física y del deporte	1577-0354	2000	Madrid	C. Virtual CC. Deporte y Universidad Autónoma Madrid/ Vicente Martínez de Haro	Trimestral
Revista jurídica de deporte y entretenimiento	1575-8923	1999	Pamplona	Aranzadi Editorial	Cuatrimestral
Selección. Revista española e iberoamericana de medicina de la educación física y el deporte	0214-8927	1989	Madrid	Federación Española de Especialistas en Medicina de la Educación Física y el Deporte	Trimestral
Tándem. Didáctica de la educación física	1577-0834	2000	Barcelona	Editorial Graó	Cuatrimestral

(*) Es continuación de **"Apuntes de medicina deportiva",** fundada en 1964. (**) Es continuación de "Derecho deportivo", fundada en 2003.

Tabla 1 (continuación). Características editoriales básicas de las revistas científico-técnicas españolas de CCAFD

acuáticas (3), fútbol (1) y danza (1). Las otras 14 revistas (43,8%) son unidisciplinares. De ellas, 5 corresponden a medicina del deporte, 1 a biomecánica, 3 a derecho deportivo, 1 a didáctica de la educación física, 3 a psicología del deporte y 1 a la psicomotricidad. Entre las revistas del inventario existen 2 que publican de manera muy marginal y casi inexistente contenidos relativos a las CCAFD. Se trata de *Aloma. Revista de psicología, ciències de l'educació i de l'esport* y A*vances en traumatología, Cirugía ortopédica, rehabilitación, medicina preventiva y deportiva*. Lo mismo venía sucediendo con la revista *Biomecánica* durante algunos años, pero desde el 2006 ha incrementado notablemente los contenidos relativos a la actividad física y el deporte.

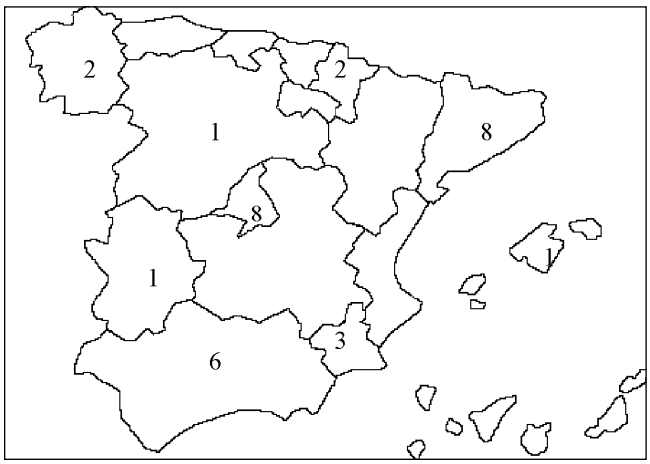


Figura 1. Distribución de la edición de las revistas por Autonomías

El análisis de las entidades editoras que figuran en la tabla 1 muestra que las sociedades científicas editan 11 revistas (34,4%), las universidades e instituciones públicas 12 (37,5%), en tanto que los editores privados publican otras 12 revistas (37,5%)[1]. Su distribución geográfica evidencia una considerable concentración de la actividad editora española, ya que ésta se distribuye entre 11 provincias y en 2 de ellas (Madrid y Barcelona) se aglutina el 50%. La provincia de Sevilla edita 4 revistas (12,5%), Murcia 3 (9,4%), La Coruña y Navarra 2 revistas cada una, mientras que Baleares, León y Cáceres editan una sola. Esta concentración en Madrid y Barcelona también se aprecia en otros áreas científicas como ciencias sociales y las humanidades (**Osca**, **Mateo**, 2003), medicina (**Ponce**, 2004) y documentación (**Delgado**, 2001). En cambio, esta tendencia desaparece cuando se analiza exclusivamente la actividad editora de las revistas electrónicas, ya que se observa una dispersión mayor por toda la geografía española (**Aguillo, Primo**, **Vázquez**, 2005). Probablemente el menor coste y la facilidad de publicar en soporte electrónico influyan en ello y permitan que instituciones pequeñas y medianas se lancen a la edición digital.

En la figura 1 se muestra la distribución de las revistas por autonomías, con 25 publicaciones (78%) editadas entre Cataluña, Madrid, Andalucía y Murcia. Este tipo de distribución está próxima a la observada en historia antigua, prehistoria y arqueología, ya que en estas mismas 4 autonomías se publica el 60,6% de las publicaciones periódicas de dichos campos (**García-Marín**, **Román**, 1998). En el conjunto de ciencias sociales y humanidades, el 62% de los títulos se editan en Madrid, Cataluña y Andalucía (**Osca**, **Mateo**, 2003). En el campo particular de la psicología son Madrid, Cataluña y Valencia las autonomías que publican el 70,6% de las publicaciones periódicas (**Osca**, **Civera**, **Tortosa**, **Quiñones**, **Peñaranda**, **López**, 2005). En todos los estudios mencionados, la Comunidad de Madrid mantiene una actividad editorial destacada, si embargo en una disciplina de las ciencias sociales como es la geografía se observa una mayor concentración de revistas en las regiones periféricas de la península (**Olcina**, **Román**, 2004).

"Durante los últimos años ha aumentado la indización de las publicaciones españolas en las bases de datos internacionales, pero todavía existe un amplio margen de mejora"

La periodicidad más frecuente es la semestral, con 12 revistas (37,5%), seguida a escasa distancia por la trimestral con 11 (34,4%). Las cuatrimestrales son 6 (18,8%), una sola revista (3,1%) tiene periodicidad bimestral y 2 (6,3%) son anuales. No hay ninguna revista que tenga una periodicidad mensual o semanal como ocurre en otros campos más antiguos y desarrollados. Estos datos indican que se trata de un tipo de periodicidad larga, más acorde con la que presentan las publicaciones periódicas de humanidades que con las de ciencia y tecnología o biomedicina. Además, sugieren la existencia de un exceso de publicaciones y/o la falta de artículos originales que, tal y como hemos dicho anteriormente, puede repercutir negativamente en la calidad de las revistas. De acuerdo con **Román** y **Gutiérrez** (2005), una revista que publica entre 2 y 4 números por año acredita a la vez un número considerable de suscriptores y una buena calidad de artículos originales. Los autores desean publicar en esa revista, lo que permite a los órganos de dirección proceder a una buena selección de los originales recibidos con el fin de promover la calidad científica de la misma. A la vez, se trata también de un índice de buena salud económica

o, al menos, de estabilidad financiera. Por el contrario, una publicación anual está casi siempre asociada a una falta de medios para editar de manera más frecuente y/o, también, a la escasez de manuscritos recibidos, lo que impide sacar más de un número por año y limita las posibilidades de selección de originales.

Una atención particular requieren los problemas de puntualidad y cumplimiento de la periodicidad de las revistas españolas de CCAFD. Así se detectan periodos sin cubrir (*Retos*) o, revistas que optan por editar con posterioridad los números atrasados como si se hubiesen publicado en el año correspondiente (*RendimientoDeportivo.com*). También se observa la publicación sistemática de números dobles para cumplir, aunque con retraso, la periodicidad propuesta (*Cuadernos de psicología del deporte*). Otras revistas que desde su comienzo y durante cierto tiempo declaran tener una determinada periodicidad cambian, sin previo aviso, a editar anualmente menos números de los previstos (*Comunicaciones técnicas*; *Fútbol. Cuadernos técnicos*; *Tándem. Didáctica de la educación física*). Asimismo, algunas revistas acumulan de forma sistemática importantes retrasos en su fecha de salida (*Fútbol. Cuadernos técnicos; Habilidad motriz*; *Revista española de educación física y deportes*). Todas estas irregularidades editoriales esconden, muy probablemente, dificultades para conseguir suficiente número de originales y suscriptores o para llevar adelante un proceso eficaz de gestión editorial. En este último caso los problemas suelen derivarse de una deficiente estructura y profesionalización de dichos procesos porque, a menudo, son asumidos por el esfuerzo aislado de un promotor individual que realiza todas las funciones editoriales (**Delgado**, 2001).

2. Difusión de las revistas

La indización en bdds selectivas, es decir, las que exigen a las publicaciones el cumplimiento de diversos criterios de calidad, se muestra en la tabla 2. Se advierte que es muy dispar, pues va desde alguna que tan sólo aparece en algún servicio documental, hasta otras que son registradas por varias bdds de prestigio.

A pesar de la importancia que tiene la indización en bdds selectivas para las propias revistas, pues entre otras ventajas aumenta de forma notable su visibilidad, sólo 11 de las 18 revistas que aparecen en la tabla 2 declaran ser indizadas por alguna bdd y las otras 7 no lo indican. Las 14 revistas restantes del total de publicaciones periódicas de CCAFD no aparecen en ninguno de los servicios documentales consultados. Al comparar estos datos con los de estudios anteriores en el mismo campo, se advierte un considerable incremento en la visibilidad puesto que en el año 2000 había indizadas 4 revistas en bdds (15,4%) (**Devís** et al., 2003), cifra que aumentó a 6 (18,8%) en 2005 (**Villamón**, **Devís**, **Valenciano**, 2005), y que ha pasado a 18 publicaciones (56,2%) indizadas en alguna bdd o servicio documental selectivo en 2007. Este incremento es similar al producido en otros campos en España como, por ejemplo, el de Ciencias de la Salud (**Vázquez, Urdín**, **Román**, 2003).

Un análisis más detallado revela que la bdd internacional *Heracles/Sportdoc*, producida en Francia, es la que más revistas españolas acoge, con 5. Existen 3 bdds (*Embase/Excerpta medica*, *PsycINFO* y *SIRC/SportDiscus*) que indizan cada una 2 revistas españolas. Otras 5 bdds selectivas y de prestigio (*Scopus*, la biblioteca virtual *SciELO*, *Pascal*, *CAB Health/CAB Abstracts*, *EbscoHost*) registran cada una de ellas los contenidos de una sola revista. Entre las bdds creadas en América Latina y que ofrecen listados de revistas clasificadas por su calidad, la brasileña *Qualis* de *Capes* califica a 4 publicaciones españolas, mientras que la hemeroteca mexicana en línea *RedALyC* registra 1 sola.

En cuanto a las bdds españolas, las del *CSIC*, es decir, *ISOC* (ciencias sociales) e *IME* (biomedicina) indizan 13 revistas entre las dos. La bdd *Ibecs* recoge 3 y, finalmente, *Psicodoc* registra 2.

Entre las bdds orientadas a ofrecer información que atiende a una serie de indicadores o de índices bibliométricos, a través de listados de revistas, cabe destacar el catálogo *Latindex* en el que se encuentran registradas **10** publicaciones periódicas del campo. DICE (Difusión y calidad editorial de las revistas españolas de humanidades y ciencias sociales y jurídicas) creada por el *Cindoc-CSIC*, registra asimismo 10 revistas. La información de esta bdd se complementa con la de otra de la misma institución, se trata de *RESH* (*Valoración integrada e índice de citas de las revistas españolas de ciencias sociales y humanidades*), en la que tan solo aparecen 3 revistas. *In-Recs* (*Índice bibliométrico de impacto de las revistas españolas de ciencias sociales*), elaborado por el grupo de investigación *EC³* de la *Universidad de Granada*, registra 5 revistas. Finalmente, otras 2 revistas electrónicas están alojadas en la plataforma digital *E-Revistas*, dentro del portal *Tecnociencia*.

La difusión en internet de las publicaciones periódicas españolas de CCAFD ha aumentado de forma destacada en los últimos años. Con la excepción de 3 revistas (*Aloma*; *RED. Revista de entrenamiento deportivo*; y *Revista de educación física*), todas las demás se han podido encontrar directamente a través del citado metabuscador *MetaCrawler* introduciendo en la casilla de búsqueda sus títulos completos o buscando en la web de los editores de las mismas. Además, al comparar la presencia de las revistas en la web con estudios anteriores, se observa claramente dicho crecimiento ya que del

Título	Bdds y servicios extranjeros	Bdds españolas	Servicios españoles
Aloma. Revista de psicologia, ciències de l'educació i de l'esport			In-Recs
Apunts. Educación Física y Deportes	Heracles/Sportdoc Catálogo Latindex Qualis (B)	ISOC	In-Recs DICE RESH
Apunts. Medicina de l'esport	Heracles/Sportdoc		
Archivos de medicina del deporte	Embase/Excerpta Medica SIRC/SportDiscus Catálogo Latindex Qualis (B)	IME Ibecs	
Biomecánica		IME	
Cuadernos de psicología del deporte	Catálogo Latindex	ISOC	In-Recs DICE
Cultura, ciencia y deporte	Catálogo Latindex Qualis (C)	ISOC	DICE
Habilidad motriz. Revista de las ciencias de la actividad física y el deporte			In-Recs DICE
Motricidad. European journal of human movement	Catálogo Latindex	ISOC	DICE
RED. Revista de entrenamiento deportivo		ISOC	
Retos. Nuevas tendencias en educación física, deporte y recreación	Catálogo Latindex	ISOC	In-Recs DICE
Revista de educación física		ISOC	
Revista de psicología del deporte	PsycINFO Ebsco Host SIRC/SportDiscus SciELO Heracles/Sportdoc Catálogo Latindex Qualis (C)	ISOC-Psic Ibecs Psicodoc	In-Recs DICE RESH
Revista internacional de ciencias del deporte. Ricyde.	PsycINFO SportDiscus/SIRC Sportdoc/ Heracles Catálogo Latindex RedALyC		DICE E-Revistas
Revista internacional de medicina y ciencias de la actividad física y del deporte	Sportdoc/Heracles RedALyC (la revista lo anuncia, pero no aparece)		E-Revistas
Revista jurídica de deporte y entretenimiento			DICE
Selección. Revista española e iberoamericana de medicina de la educación física y el deporte	Scopus Embase/Excerpta Medica Pascal CAB Health/CAB Abstracts Catálogo Latindex	IME Ibecs Psicodoc	
Tándem. Didáctica de la educación física	Catálogo Latindex	ISOC	In-Recs DICE RESH

Tabla 2. Bases de Datos y Servicios documentales selectivos que indizan a las revistas

30,8% del año 2000 (**Devís** et al., 2003) se ha pasado al 90,6% en 2007. Estos últimos datos son similares a los obtenidos por las revistas jurídicas (**Gómez del Pulgar**, 2006) y las de ciencias de la salud (**Vázquez, Urdín**, **Román**, 2003).

Por lo que respecta al tipo de acceso vía internet, se observa que casi la mitad (46,9%) ofrece la consulta a través de una web propia e independiente, muy por encima del 3,7% de las revistas jurídicas españolas (**Gómez del Pulgar**, 2006). Para el resto se hace a través de un enlace (37,5%) y sólo 2 son accesibles consultando el catálogo de publicaciones correspondiente en la web de su editor (ver tabla 3).

Quizá el aspecto de más interés para los lectores, en relación con la difusión de las revistas en internet, es el contenido disponible. Es muy variado, y en la tabla 3 se ha intentado categorizarlo a través de siete posibilidades. Así, por ejemplo, 26 revistas ofrecen una referencia básica, aunque en ocasiones es tan pobre que en ella no aparecen datos tan importantes como el ISSN o la periodicidad, tal y como sucede con las revistas *Agua y gestión; Cairon; Comunicaciones técnicas* y *Revista andaluza de derecho del deporte*. Un grupo de 17 revistas ofrecen el acceso a todos los sumarios y otras 13 presentan una versión íntegra de sus contenidos (5 de pago y 8 gratuitas). Las publicaciones que ofrecen la versión íntegra en internet suponen un 40,6%, un porcentaje nada despreciable. También existen otras 2 revistas (*Avances* y *Revista española de educación física y deportes*) que ofrecen algún número completo escaneado en *pdf*, y sin embargo, paradójicamente, no presentan ninguna página con su información básica.

		Nº	%
Tipo de acceso	Dirección propia	15	46,9
	Desde web del editor	12	37,5
	Desde catálogo	2	6,2
Contenidos	Referencia básica	26	81,2
	Último sumario	19	59,3
	Varios sumarios	3	9,3
	Todos los sumarios	17	53,1
	Algún trabajo completo	6	18,7
	Versión íntegra	13	40,6
	Información complementaria	16	50,0
Servicios	Suscripción por internet	9	28,1
	Correo electrónico de contacto	14	43,7
	Buscador	1	3,1

Tabla 3. Tipo de acceso, contenidos y servicios ofertados en internet por las revistas

"El 81,2% de las revistas están presentes en internet al menos con una referencia básica, algo más de la mitad permite el acceso a todos los sumarios y una cuarta parte ofrece el texto completo gratuito"

Por último, los servicios que proporcionan a través de internet, un valor añadido a la edición tradicional en papel, se limita a una simple dirección de correo electrónico para contactar con los editores en 14 revistas de CCAFD. La posibilidad de realizar la suscripción mediante un formulario elaborado al efecto lo ofrecen otras 9, mientras que sólo 1 de ellas (*Apunts. Medicina de l'esport*) incluye un sistema de búsqueda.

Conclusiones

El estudio sobre las características y difusión de las revistas científico-técnicas españolas de CCAFD ofrece las siguientes conclusiones:

– Existe un total de 32 publicaciones periódicas en el año 2007, 6 títulos más que en el año 2000.

– Las revistas editadas exclusivamente en soporte papel son el 28,1%, frente al 56,2% que lo hacen en papel y electrónicamente y el 15,6% sólo electrónicamente.

– El 65,6% de las revistas se han fundado con posterioridad a 1991 y cerca de la mitad de ellas en los últimos diez años.

– Existen más revistas multidisciplinares (56,2%) que unidisciplinares (43,8%).

– Las asociaciones científicas y las universidades e instituciones públicas editan más de dos terceras partes de las revistas. Las instituciones privadas, principalmente editoriales comerciales, publican otra tercera parte.

– La mitad de las revistas se editan entre las provincias de Barcelona y Madrid. Las autonomías de Cataluña, Madrid, Andalucía y Murcia editan más del 78% de las publicaciones.

– Aproximadamente la mitad de las revistas españolas de CCAFD tiene una periodicidad semestral y anual.

– Existen 11 revistas (34,4%) que declaran estar indizadas por alguna bdd, aunque otras 7 (21,9%) también lo están y no lo indican.

– Las bases de datos internacionales de tipo selectivo indizan 12 revistas (37,5%). Hay 1 revista que está presente en 7, y 2 en 5 bases de datos.

– El 81,2% de las revistas está presentes en internet al menos con una referencia básica. Los contenidos que ofrecen son muy variados, destacando que el 53,1% permite el acceso a todos los sumarios y el 25% al texto completo gratuito.

Como puede observarse, las características editoriales básicas indican que las CCAFD es un campo joven y poco consolidado académicamente. El número de títulos sigue creciendo, pero la mitad de las revistas son de una periodicidad semestral y anual, lo cual puede indicar demasiadas revistas y/o poca cantidad de artículos originales. Además, el predominio de revistas multidisciplinares indica una baja especialización. Una oferta excesiva de títulos influye negativamente en la calidad de las revistas al dispersarse también el número de artículos originales potenciales, disminuyendo la solidez y posicionamiento internacional de las revistas españolas de CCAFD. La distribución geográfica de la edición es similar al de otras disciplinas de ciencias sociales y humanas. La indización en las bases de datos internacionales ha mejorado en los últimos años, pero podría incrementarse bastante más. Asimismo, se observa una marcada tendencia a la presencia de las revistas en internet, aunque los servicios que ofrecen son todavía muy básicos.

Nota

1. La suma total de 35 revistas, en lugar de las 32 del inventario, es debida a que en 3 casos la edición es asumida por 2 entidades/actores a la vez.

Reconocimiento

Agradecemos al *Ministerio de Educación* la ayuda recibida por el proyecto de investigación *SEJ2004-03996/EDUC* con la que se ha financiado este trabajo

Bibliografía

Abad, María-Francisca; **González, Aurora**; **Martínez, Celeste**. "Características de las revistas médicas españolas. 2004". En: *El profesional de la información*, 2005, septiembre-octubre, v. 14, n. 5, pp. 380-390.

Aguillo, Isidro; **Primo, Elena**; **Vázquez, Manuela**. "Evaluación de las revistas electrónicas de ciencias de la salud, editadas en España frente a los criterios de calidad del sistema Latindex". En: *XI Jornadas de información y documentación en ciencias de la salud*, Terrassa, 2005. Consultado en: 13/04/2006.
http://www.jornadasbibliosalud.net/c34.pdf

Aleixandre-Benavent, Rafael; **Valderrama-Zurián, Juan-Carlos**; **González-Alcaide, Gregorio**. "El factor de impacto de las revistas científicas: limitaciones e indicadores alternativos". En: *El profesional de la información*, 2007, enero-febrero, v. 16, n. 1, pp. 4-11.

Aliaga, Francisco M.; **Suárez, Jesús**. Tendencias actuales en la edición de revistas electrónicas: nueva etapa en *Relieve*. En: *Revista electrónica de investigación y evaluación educativa,* 2002, v. 8, n. 1. Consulta en: 7/08/2004.
http://www.uv.es/RELIEVE/v8n1/RELIEVEv8n1_0.htm

Aquesolo, José. "Apuntes para una historia de la documentación deportiva". En: *Revista general de información y documentación*, 2000, v. 10, n. 1, pp. 31-67.

Delgado-López-Cózar, Emilio. "Las revistas españolas de ciencias de la documentación: productos manifiestamente mejorables". En: *El profesional de la información,* 2001, diciembre, v. 10, n. 12, pp. 46-56.

Devís, José; **Antolín, Luis**; **Villamón, Miguel**; **Moreno, Alberto**; **Valenciano, Javier**. "Las revistas científico-técnicas españolas de las ciencias de la actividad física y el deporte: inventario y análisis de la calidad de contenido y difusión". En: *Revista española de documentación científica*, 2003, v. 26, n. 2, pp. 177-190.

Devís, José; **Villamón, Miguel**; **Antolín, Luis**; **Valenciano, Javier**; **Moreno, Alberto**. "Las revistas científico-técnicas españolas de ciencias de la actividad física y el deporte: adecuación a las normas ISO y grado de normalización". En: *Ciência da informação,* 2004, v. 33, n. 1, pp. 38-47.

García-Marín, Ángeles; **Román, Adelaida**. "Las publicaciones periódicas de historia antigua, prehistoria y arqueología: difusión internacional". En: *Trabajos de prehistoria*, 1998, v. 55, n. 1, pp. 139-146.

Giménez, Elea; **Gómez, Isabel**; **Vázquez, Manuela**. "Difusión nacional e internacional de las revistas científicas". En: **Román, Adelaida** (coord.). *La edición de revistas científicas. Guía de buenos usos*. Madrid: CINDOC, 2001, pp. 35-46. ISBN 84-00-07916-7.

Gómez-del-Pulgar, Gloria. "Difusión por internet de las revistas jurídicas españolas de edición impresa". En: *Revista española de documentación científica,* 2006, v. 29, n. 2, pp. 258-285.

Hopkins, Will G. (2004). "Impact factor of journals in sport and exercise science, 2000-2003". *Sportscience*, n. 8, pp. 12-19. Consulta en: 7/07/2006.
http://sportsci.org/jour/04/wghif.htm

Hopkins, Will G. (2005). "Impact factor of journals in sport and exercise science, 2004". *Sportscience*, n. 9, pp. 14-16. Consulta en: 7/07/2006.
http://www.sportsci.org/jour/05/wghif.htm

Hopkins, Will G. (2007). "The Tour de Journals 2007: Impact Factors in Exercise and Sport". *Sportscience*, n. 11, pp. 9-11. Consulta en: 13/09/2007.
http://www.sportsci.org/j2007/wghif.htm

Krauskopf, Manuel; **Vera, María-Inés**. "Las revistas latinoamericanas de corriente principal: indicadores y estrategias para su consolidación". En: *Interciencia*, 1995, mayo-junio, v. 20, n. 3, pp. 144-148.

Olcina, Jorge; **Román, Adelaida.** "Las revistas españolas de geografía: cambios y adaptación a los criterios editoriales de calidad". En: García Ramón, María Dolores (eds.) Proceedings. *La geografía española ante los retos de la sociedad actual: aportación española al XXX Congreso de la Unión Geográfica Internacional*. Glasgow (UK), 2004, pp. 145-179. Consulta en: 23/03/2005.
http://eprints.rclis.org/archive/00002884/.

Osca, Julia; **Mateo, María Elena**. "Difusión de las revistas españolas de ciencias sociales y humanidades. Acercamiento bibliométrico". En: *Revista general de información y documentación*, 2003, v. 13, n. 1, pp. 115-132.

Osca, Julia; **Civera, Cristina**; **Tortosa, Francisco**; **Quiñones, Elena**; **Peñaranda, María**; **López, Juan José**. "Difusión de las revistas españolas de psicología en bases de datos nacionales e internacionales". En: *Anales de documentación*, 2005, n. 8, pp. 165-186.

Ponce, Concepción. *Análisis de la circulación de las revistas biomédicas españolas en bases de datos nacionales e internacionales* (Tesis doctoral). Valencia: Universitat de València, 2004.

Román, Adelaida; **Vázquez, Manuela**; **Urdín, Carmen**. "Los criterios de calidad editorial Latindex en el marco de la evaluación de las revistas españolas de humanidades y ciencias sociales". En: *Revista española de documentación científica*, 2002, v. 25, n. 3, pp. 286-307.

Román, Adelaida; **Gutiérrez, Beatriz**. "Étude sur les revues espagnoles en sciences humaines et sociales". En: **Minon, Marc**; **Chartron, Ghislaine** (coords.). *État des lieux comparatif de l'offre de revues SHS*, *France-Espagne-Italie*. Rapport pour le Ministère français de la recherche, 2005, juin, pp. 27-46. Consulta en: 2/08/2006. *http://archivesic.ccsd.cnrs.fr/docs/00/06/26/64/PDF/sic_00001561.pdf*

Urdín, Carmen; **Vázquez, Manuela**; **Román, Adelaida**. "Los criterios de calidad editorial Latindex en el marco de evaluación de las revistas españolas de ciencia y tecnología". En: *Revista española de documentación científica*, 2003, v. 26, n. 1, pp. 56-73.

Vázquez, Manuela; **Urdín, Carmen**; **Román, Adelaida**. "Las revistas españolas de ciencias de la salud frente a los criterios de calidad editorial Latindex". En: *Revista española de documentación científica*, 2003, v. 26, n. 4, pp. 418-432.

Villamón, Miguel; **Devís, José**; **Valenciano, Javier**. "Análisis de la visibilidad de las revistas científico-técnicas españolas de ciencias de la actividad física y el deporte". En: *Revista de psicología del deporte*, 2005, v. 14, n. 2, pp. 253-267.

Villamón, Miguel; **Devís, José**; **Valenciano, Javier**. "Análisis de las 'Instrucciones para autores' de las revistas españolas de ciencias de la actividad física y el deporte". En: *Motricidad*, 2006, n. 16, pp. 133-150.

Miguel Villamón, ***José Devís***, ***Alexandra Valencia*** *y* ***Javier Valenciano***. *Facultad de Ciencias de la Actividad Física y el Deporte, Universitat de València. Departamento de Educación Física y Deportiva. C/ Gascó Oliag, 3. 46010 Valencia.*
miguel.villamon@uv.es
jose.devis@uv.es
vapea@alumni.uv.es
francisco.valenciano@uv.es

EARLY BIRD AND ASSOCIATION MEMBERSHIP DISCOUNTS AVAILABLE

online information 2007

WWW.ONLINE-INFORMATION.CO.UK/CONFERENCE

Applying Web 2.0: Innovation, Impact and Implementation

4–6 December
Olympia Grand Hall, London, UK

THE NO.1 CONFERENCE FOR THE INFORMATION WORLD

Opening keynote speaker
Jimmy Wales, Founder, **Wikipedia & Wikia**

- **Learn** from over 100 international information leaders including: **Euan Semple**, Social Computing Expert, **Greg Notess**, Montana State University, **Stephen Abram**, SirsiDynix and **Gunnar Sahlin**, National Library of Sweden
- **Discover** what Web 2.0 means for you
- **Experience** Web 2.0 in practice with case studies from: Vodafone, BT, IBM, Drugscope, University of Sheffield, Mozilla Europe and more
- **Gain** new skills and competencies for information professional 2.0
- **Hear** from experts at more than 30 sessions covering all aspects of the information industry over 3 days
- **Network** with your peers from over 45 countries
- **Develop** your **'Web 2.0'**, **'Web Search'**, **'Research & Competitive Intelligence'** skills and knowledge in the workshops on Monday 3 December
- **See** the latest industry developments at the **2 co-located free exhibitions** of over 250 online content and information management solutions providers

MEDIA PARTNER: iWR

PLATINUM CONFERENCE SPONSOR: INNODATA ISOGEN

SUPPORTED BY: aiip, ALPSP, Aslib, CiG, cilip, CILIP, dfid, EUSIDIC, FreePint, The Information Architecture Institute, records management society, SLA, SPIN, UKeiG

An incisivemedia event

View full programme information and book your place at www.online-information.co.uk/conference

Evolución y uso de los lenguajes controlados en documentación informativa

Por **Lourdes Castillo** y **Alejandro de-la-Cueva**

Resumen: *La documentación periodística adolece de instrumentos adecuados de clasificación e indización de la información, a excepción del Subject Reference System del Internacional Press Telecommunications Council (IPTC), todavía en estado incipiente de desarrollo. Se hace una revisión de las contribuciones más relevantes sobre la clasificación e indización de noticias en los medios de comunicación, tanto españoles como internacionales. También se estudia la elaboración y uso de vocabularios controlados para el tratamiento de la información de actualidad y sobre todo de tesauros especializados. Por último, se destacan algunas características específicas de la documentación periodística que condicionan la utilización de tesauros para la indización y recuperación de información de actualidad.*

Palabras clave: *Documentación periodística, Lenguajes controlados, Tesauros, Clasificaciones, Indización de prensa, Servicios de documentación de medios de comunicación.*

Title: Evolution and use of controlled languages in news documentation

Abstract: *News documentation lacks suitable instruments of classification and indexing, the only exception being the Subject Reference System of the International Press Telecommunications Council (IPTC), which is not yet fully developed. The most relevant contributions to classifying and indexing news in Spanish-language media, both in Spain and internationally, is reviewed. We also studied the development and use of controlled vocabularies for processing news information, primarily specialized thesauri. Finally, news documentation characteristics specific to indexing and information retrieval for print news media are described in detail.*

Lourdes Castillo *es doctora en geografía e historia por la Universitat de València. Ha impartido clases en la diplomatura de biblioteconomía y documentación de la Universitat de Valencia y es documentalista en la Unidad de Documentación de RTVV.*

Alejandro de la Cueva*, doctor en cirugía y medicina (1987) por la Universidad de Valencia. Especialista en documentación médica por la Universidad de Valencia (1990). Profesor titular de biblioteconomía y documentación del Departamento de Historia de la Ciencia y Documentación de la Universidad de Valencia, desde 1988. Analista de la base de datos Índice Médico Español (IME) desde 1973 a 1992. Vicedirector de la misma entre 1979 y 1989.*

Keywords: *News documentation, Controlled languages, Thesauri, Classifications, News indexing, News reference services.*

Castillo, Lourdes; Cueva, Alejandro de la. "Evolución y uso de los lenguajes controlados en documentación informativa". En: *El profesional de la información,* 2007, noviembre-diciembre, v. 16, n. 6, pp. 617-626.

DOI: 10.3145/epi.2007.nov.08

Introducción

EN ESPAÑA SE EDITAN 95 PERIÓDICOS de los cuales 66 tienen edición electrónica, es decir un 69% del total[1]. En el mundo las ediciones digitales de la prensa en lengua española constituyen por su volumen el segundo segmento lingüístico aunque distantes en número de los 1.236 diarios en inglés (Díaz Nosty, 1999).

Un archivo periodístico de tamaño medio, como el de *Radiotelevisión Valenciana* (*RTVV*), selecciona un total de 289.631 noticias desde 1998 hasta 2005, con un crecimiento medio de más de 36.000 noticias al año. Un volumen mayor, de aproximadamente 139.404 noticias correspondientes a los años 2003 a 2005, contiene la base de datos de prensa de *Euskal Telebista*. Por otra parte, los servicios de seguimiento de prensa pueden almacenar volúmenes ingentes de noticias. Es el caso del servicio de información de prensa digital *iMente*, que diariamente captura unas 115.000 noticias diarias (**Guallar**, 2006).

El acceso inmediato a la información periodística y la interactividad son dos de las principales características que diferencian a los diarios digitales, pero no las únicas. La posibilidad de actualización y corrección continua; la sencillez para el usuario de copiar, editar y archivar los textos íntegros; el acceso a los números atrasados y a colecciones enteras de periódicos son otras de las posibilidades que ofrece la prensa distribuida a través de internet. Aparentemente el acceso casi generalizado a las telecomunicaciones parece solucionar el problema de la búsqueda de información; sin embargo los problemas en la recuperación de información en cuanto a calidad y pertinencia persisten y es necesario seguir investigando, además de estudiar cómo adaptar las herramientas tradicionales a los nuevos entornos y usuarios, los que se han dado en llamar "usuarios casuales", con habilidades de recuperación limitadas.

El control de vocabulario es una realidad en documentación científica y en otras áreas como la documentación administrativa o la legislativa y jurídica. Esa normalización se refleja en herramientas de control que se aplican en grandes sistemas como *PubMed* o *FSTA*. Algunos ejemplos de estas herramientas terminológicas son tesauros como el *MeSH*, el *Agrovoc*, el macrotesauro de la *OCDE*, *Eurovoc*, *Hasset*, *Sosig*, *Glin*, *Spines*, *Eric*, *Envoc*, y muchos más.

Frente a esta normalización contrasta la ausencia en documentación periodística de instrumentos generales consensuados respecto a la indización y clasificación de documentos. Cada medio de comunicación utiliza su propio lenguaje controlado, elaborado independientemente; y en muchas ocasiones se trata de vocabularios de bajo nivel como clasificaciones o listas de autoridades. A esto hay que añadir una tendencia que no ha beneficiado nada al desarrollo de los lenguajes controlados en documentación periodística: el empleo del lenguaje natural en la recuperación de información de actualidad.

"Cada medio de comunicación utiliza su propio lenguaje controlado, elaborado independientemente; y en muchas ocasiones se trata de vocabularios de bajo nivel, como simples listas"

Tan sólo el *Subject Reference System*[2], la clasificación temática desarrollada por el *Consejo Internacional de Telecomunicaciones en Prensa* (*Internacional Press Telecommunications Council*, *IPTC*), puede representar este instrumento consensuado, pero todavía poco desarrollado.

El presente trabajo tiene como objetivo hacer una revisión de estudios relevantes sobre la evolución de la clasificación e indización de los documentos de prensa y la elaboración y utilización de vocabularios controlados, especialmente del tesauro, en documentación informativa. También se resaltan algunas de las causas que se señalan en la literatura como condicionantes de la escasa utilización de los tesauros en este medio.

Estudios sobre la documentación y la indización de la prensa

El decano de la indización de prensa es **William Frederick Poole**, que en 1848 publicó *Index to subjects treated in the reviews and other periodicals.* En 1853 mejoró este primer esfuerzo con el título *Index to periodical literature*, obra considerada precursora del *The New York times index* (**Semonche**, 1993). En 1902 **Henry B. Wheatley** publicó *How to make an index*, que es una guía sobre el proceso de indización (**Semonche**, 1993). Hasta 1909 los estudios sobre los servicios de documentación periodística fueron escasos y esporádicos. Quizás el primer artículo sobre la actividad documental informativa fue el de **I. D. Marshall**, "Article on methods in newspaper libraries", publicado en una de las primeras revistas de periodismo, *Newspaperdom*, en su primer número de 1892. Este texto hace las primeras observaciones sobre la labor de clasificación y archivo de los materiales de referencia. En 1893 **Bardwell** se refirió también a las tareas de clasificación y archivo de recortes en el artículo *Scrapbooks*, *Clippings*, *etc*., publicado en *Library journal* (**Galdón**, 1986).

González-Quesada (1995) explica cómo se constituyó en 1923, dentro de la *Special Libraries Association*, un *Grupo de Periodismo* dedicado a analizar y sistematizar el funcionamiento de los servicios de documentación en las empresas periodísticas. La revista *Special libraries* permitió hacer públicos los estudios iniciales sobre los servicios de documentación periodística y contribuyó decisivamente a dar a conocer los métodos empleados en la labor de clasificación y archivo de diversos rotativos norteamericanos. **Hansen** y **Ward** (1991) publicaron un trabajo sobre el uso de bases de datos electrónicas y otras tecnologías de la información en 105 diarios de gran tirada en 1990. Por su parte **Hegg** (1991) hizo un estudio comparativo sobre el almacenamiento y los métodos de recuperación automatizados en pequeños diarios de EUA. El *Grupo de Periodismo* constituyó además un comité para elaborar una clasificación estándar, pero no lograron su propó-

sito. De hecho, en 1933 no había dos centros de documentación periodísticos con el mismo sistema de clasificación, aunque hubiese coincidencias parciales. **Galdón** (1994) enumera los criterios que seguían la mayoría: elaborar una clasificación específica para recortes periodísticos; acotarla en apartados de temas, personas y países; fijar encabezamientos y subdivisiones simples y fáciles de recordar y permitir la codificación por varias entradas.

En su capítulo sobre indización de prensa de 1993, **Semonche** destaca las aportaciones de **Desmond** y **Friedman**. **Robert W. Desmond** publicó en 1933 *Newspaper reference methods*. Este libro analiza y valora los aspectos de estructura y método, a la vez que hace hincapié en el tratamiento de los problemas derivados de la clasificación. **Harry Friedman** escribió en 1942 *Newspaper indexing*, otra obra sobre indización de prensa.

La *American Newspaper Publisher Association* (*ANPA*) publicó en 1974 *Guidelines for newspaper libraries*, reeditada en 1976. Las *Guidelines* constan de 16 capítulos sobre otros tantos aspectos que, según la *Newspaper Division* de la *Special Libraries Association*, se deben tener en cuenta al plantear los servicios de documentación periodística. Uno de estos capítulos está dedicado a la indización y expone los diferentes modos de realizar índices y la forma de usarlos.

En el ámbito europeo la obra más destacable a nivel teórico es la de **Geoffrey Whatmore**, que en 1964 publicó *News information: the organization of press cuttings in the libraries of newspaper and broadcasting services*, traducido en España en 1970 con el título *La documentación de la noticia: organización y métodos de trabajo para archivos de referencia de periódicos y agencias*. Este libro consta de 17 capítulos y en dos de ellos se hace referencia a los índices, pero siempre hablando de sistemas manuales. Expone los sistemas de índices de varios periódicos (*Financial times*, *The guardian*, *Daily mail*, *Daile sketch*, *Evening news*, *Glasgow herald*) en forma de tarjetas con marcadores de colores, tiras superpuestas con márgenes inferiores visibles, en volúmenes y hojas intercambiables.

"A nivel internacional, el tesauro de prensa más destacado por todos los autores es el empleado en la base de datos del diario *The New York times*"

En 1973, con ocasión del primer congreso sobre servicios de documentación periodística organizado en Gran Bretaña por la *Association of Special Libraries and Information Bureaux* (*Aslib*), **Whatmore** expuso, bajo el título *Classification for news Libraries*, los problemas de clasificación específicos de estos servicios. El mismo autor publicó en 1978 el libro *The modern news library*, en el que pone al día los métodos prácticos de archivo y organización de los servicios de documentación, pormenorizando las soluciones a las posibles dificultades prácticas y aportando una lista de pautas a seguir en las tareas de selección y clasificación. En 1979 se publicaron las *Normas para la presentación de índices analíticos en centros de documentación y archivos de publicaciones periódicas,* preparadas años antes por **Justo García-Morales**, director del *Centro Nacional del Tesoro Documental y Bibliográfico*. **García-Morales** (1979) afirma que la dificultad de organizar un centro de documentación de prensa estriba en la universal amplitud de las materias, en el tiempo y en el espacio geográfico que reflejan. Aconseja que el sistema adoptado sea ante todo práctico y rápido, considerando que el objeto de la elaboración de índices es facilitar la consulta urgente y la localización de cualquier dato o curiosidad, contenido en un diario o publicación de índole informativa.

En 1991 **Hans H. Wellisch** publicó *Indexing from A to Z*, en el que ofrece algunos consejos y consideraciones sobre la indización de la prensa. Dos años más tarde **Barbara P. Semonche** editó el libro *News media libraries: a management handbook*, en cuyo capítulo 19, "Newspaper indexing policies and procedures", escrito por la propia autora, repasa la historia y los tipos de índices de periódicos, ofreciendo pautas generales para indizar un periódico.

A nivel internacional, el tesauro de prensa más destacado por todos los autores es el empleado en la base de datos del diario *The New York times*, que data de 1971 y puede considerarse como el primero de los de información de actualidad. Incorpora más de 700.000 descriptores e incluye una relación alfabética de nombres de personas.

Alan R. Greengrass (1983), analiza el funcionamiento del departamento de indización de este diarioasí como la estructura del tesauro empleado y del índice impreso; describe el proceso de toma de decisiones para la incorporación de nuevos términos y los problemas de la elaboración del tesauro. Además estudia su base de datos de texto íntegro proporcionando ejemplos de resúmenes y de páginas del tesauro.

El tesauro del diario *The New York times* ha sido estudiado por varios autores. **Pastor** (1992) expone sus características principales y la

estructura y relaciones semánticas. También **Caridad** (1980) y **Coll-Vinent** (1982) han publicado trabajos que proporcionan información sobre el mismo.

Milstead (1983) estudia los índices de prensa de diarios norteamericanos y considera que, de entre los publicados, los más detallados son los de *The New York times index*, *The London times index* y *The official Washington post index*. Cada uno de ellos emplea su propio vocabulario controlado, con sus múltiples referencias cruzadas y sus subdivisiones jerárquicas. El *National newspaper index* proporciona sólo una lista de términos, basada en la *Library of Congress subject headings*. En este mismo artículo **Milstead** describe el caso concreto del índice del *The Washington post*, la especialización temática de los indizadores, el procedimiento de indización, el empleo de referencias cruzadas y subencabezamientos, la revisión mensual y anual de los índices, la consideración como esencial del vocabulario controlado, la inadecuación de otros índices y de las listas de encabezamientos a su nivel de especificidad y la estructura del índice.

Martínez-Pestaña (1986) describe 54 bases de datos periodísticas y solamente en tres de ellas se hace referencia a la indización: la del *The New York times*, en la que como ya se ha visto, se especifica la indización con vocabulario controlado, la de *The lexicon herald leader*, que incluye descriptores y la de *Ontap magazine index*.

Pastor (1992) detalla en unas tablas 51 bases de datos de prensa internacionales aparecidas entre 1971 y 1991 y especifica entre otros datos el tipo de lenguaje documental utilizado en el análisis y recuperación de la información. Sólo nueve de ellas, un 17%, disponen de un tesauro de información de prensa para el análisis y recuperación de la información: *The New York times; Gruner & jahr, Politiken dagbladet; St. Louis post dispatch; Documentation française, BIPA; Ringier & Co; Aftonbladet; Edi 7* y *Le monde*. Esta autora expone las principales características de los tesauros de *The New York times* y *Gruner & jahr* debido a que, durante mucho tiempo, han sido los únicos vocabularios adecuadamente estructurados y desarrollados en el área de la documentación periodística.

El tesauro de la base de datos periodística de la editorial alemana *Gruner & jahr* contiene un total de 5.000 descriptores relacionados entre ellos por referencias cruzadas. Se encuentran integrados en 32 grandes facetas o campos temáticos, subdivididos a su vez en varios apartados. Las relaciones que recoge son las de equivalencia y jerarquía.

En el *Segon seminari: l'experiencia multimedia*, **Fuentes** (1994) anota el empleo de descriptores en las bases de datos de *Le monde* y de *Bayard-presse*. También **Coll-Vinent** (1978) estudió el centro de documentación de *Le monde*.

El tesauro de *Le monde diplomatique*[3], pese a denominarse así, es más bien una clasificación o una lista de términos autorizados de temas amplios y no muy numerosos. Incluye también ámbitos geográficos.

González-Quesada (1995), en un estudio sobre la evolución histórica de la documentación periodística, comenta que en los años 70 cuarenta diarios norteamericanos y una docena de europeos –entre los que se hallaban los británicos *The times* y *The guardian*, el francés *Le monde*, el italiano *Corriere de la sera* y el alemán *Frankfurter allgemeine zeitung*– utilizaban índices como guía de los contenidos de la publicación. El mismo autor afirma que la *BBC* cuenta con un sistema de clasificación propio, el *Schedule of subject headings*.

En mayo de 1998 se inició el proyecto de la Unión Europea *Laurin* con el objetivo de desarrollar un modelo genérico para la digitalización de recortes de prensa. Gestionado y coordinado por el *Department of German Language and Literature* de la *Universidad de Innsbruck*, es poseedor de una de las más amplias colecciones de recortes de prensa de literatura y crítica procedentes de prensa de Austria, Alemania y Suiza. Otras siete colecciones de recortes de prensa sobre cultura, política y economía de otros tantos países europeos se unieron al proyecto. Todos esos casos junto a los nuevos que se añaden, forman la base de datos *Laurin*. Los recortes están enlazados con el tesauro multilingüe *Laurin* que está organizado por conceptos, además incluye los topónimos del *Getty thesaurus geographic names* y los de la *Nomenclature of territorial units for statistics.* (**Mühlberger**, 1999; **Calvanese**, et al., 2001).

El diario *Les echos* clasifica las noticias de acuerdo con unas listas predefinidas de materias y ámbitos geográficos. *Los Angeles times* admite en las búsquedas online emplear los descriptores que asignan los documentalistas (**Jiménez** et al., 2000).

El *Web thesaurus compendium*[4] no incluye ningún tesauro especializado en medios de comunicación o especialmente válido para la información de actualidad.

En 1999 el *IPTC* desarrolló la *Subject reference system* para permitir a los proveedores de información (principalmente agencias) acceder a un lenguaje independiente codificado para describir el contenido de las noticias y para facilitar su intercambio. El sistema se explica en la *Information interchange model guideline 3* (*IIMG3*)[5]. Se trata de un sistema clasificatorio jerárquico y consta de 17 niveles temáticos principales que, a su vez, se sub-

"Los estudios sobre lenguajes controlados en medios españoles evidencian una escasez en la elaboración y uso de tesauros debidamente estructurados"

dividen en términos relacionados jerárquicamente. Hay que recordar que en España el grupo *Prisacom* lo adaptó en 2001 para utilizarlo como sistema de indización y de recuperación para su hemeroteca digital.

Los vocabularios controlados en los medios de comunicación españoles

Los estudios sobre el empleo de lenguajes controlados en medios de comunicación españoles evidencian un uso predominante de glosarios y sistemas clasificatorios, además de una escasez en la elaboración y uso de tesauros debidamente estructurados.

Martín-Muñoz y **López-Pavillard** (1995) citan algunos de los tesauros, temáticamente independientes (geográfico, de deportes, de agricultura, de animales, de ciencia y técnica) que se emplean en las bases de datos de documentación audiovisual de *Radiotelevisión Española* (*RTVE*) y que fueron creados y desarrollados por el propio centro de documentación a lo largo de su actividad. Para la base de datos de documentación escrita de *RTVE*, *Basinfa* (*Base de Información de Actualidad*) se consideró imprescindible contar con dos tesauros; uno temático y otro geográfico. Ante la imposibilidad de elaborar un tesauro temático propio se optó por realizar una adaptación del elaborado por la *Unesco*.

En cuanto a *Telecinco*, para el análisis de la información audiovisual se emplea un tesauro que incluye 3.000 descriptores controlados y cerca de 1.000 no descriptores con sus correspondientes reenvíos. **Valle-Gastaminza**, **García-Jiménez** y un equipo de colaboradores (2001) analizan el estudio, la construcción y su puesta en funcionamiento.

Llobet y **Pañella** (1988) describieron el proceso de la elaboración del tesauro para la base de datos de imágenes de *Televisió de Catalunya* (*TV3*). Elaborado con el programa *Mistral*[6], se encuentra estructurado en dos partes: la primera incluye nombres propios de personas y entidades (diccionario de autoridades) con relaciones de sinonimia y notas explicativas. La segunda está dedicada a nombres comunes y geográficos que incluyen relaciones jerárquicas (genéricos y específicos), asociativas (términos relacionados) y de sinonimia. Se emplea también para la descripción de imágenes en la *Unidad de Documentación* de *Radiotelevisión Valenciana*. El *Departamento de documentació* de *Televisió de Catalunya* ha sido analizado en la tesis de **Codina** (1996).

El *Centro de documentación y archivo* de *Euskal Telebista* (*ETB*) utiliza para la recuperación temática de su base de datos de imágenes, un tesauro elaborado por su equipo de documentalistas a partir del de *TV3*. Consta de 12.000 términos en total: 3.500 son descriptores temáticos, 2.500 geográficos y 6.000 identificadores o descriptores onomásticos. Para la base de datos de noticias de prensa utiliza una lista de materias de elaboración propia realizada a partir de diversos listados y tesauros.

Hay que señalar que las bases de datos de imágenes de televisión solamente disponen de representaciones textuales de las imágenes. Las de prensa escrita sin embargo, disponen ya de los textos íntegros y por lo tanto de la posibilidad de utilizar todos los términos de los artículos. Esto significa que además de por la naturaleza de la tipología documental, las bases de datos de imágenes audiovisuales requieren unos vocabularios controlados menos conceptuales y no son totalmente válidos para la recuperación de información de noticias de prensa. Los vocabularios controlados para la descripción de imágenes de televisión están sometidos a un esquema de datos mínimos de representación recomendado internacionalmente, la *minimum data list* de la *FIAT/IFTA*[7], que contiene recomendaciones sobre la representación de contenidos.

Por otra parte, en los centros de documentación de diarios españoles la situación es similar a la de los medios audiovisuales. El servicio de documentación del diario *El correo español* fue analizado en la tesis de **Pastor-Ruiz** (1992). En la tercera parte de este trabajo se describe el lenguaje documental empleado en las bases de datos de este periódico. Se trata de un vocabulario controlado compuesto por dos listas alfabéticas construidas con las palabras clave registradas, durante la fase de indización, en los campos de materias y de descriptores. **Martín** (1994) también describe el centro de documentación y las bases de datos de este diario. El autor hace referencia al mantenimiento de cuatro listas: descriptores, personas, lugares y modos, que comparten las bases gráficas y de prensa. Comenta además que la falta de modelos y la inexistencia de herramientas de control de vocabulario propiciaron un crecimiento exagerado del número de términos, facilitado también por una filosofía de precoordinación muy acusada y redundante.

En la descripción realizada por **Aguado** (1995) del sistema de ar-

chivo y documentación del *Grupo Prensa Española*[8] (diario *ABC* y revista *Blanco y negro*) se mencionan los campos onomásticos, geográficos y temáticos en los que se emplean tablas de validación y tesauros especializados adaptados a las necesidades propias de *Prensa española*. No se dan más detalles de estos tesauros, sin embargo, se menciona en el mismo artículo la elaboración de uno de ellos, así como su aplicación a los índices como objetivo prioritario.

En una descripción del diseño y creación de la base de datos documental del *Grupo Godó*[9], **Salmurri** et al. (2002), mencionan estar desarrollando herramientas (listas de validación y tesauro) para la descripción documental y de contenidos.

En cuanto a *El país*, el diario español de mayor tirada, utiliza una extensa clasificación jerárquica. A partir de su base de datos documental se generaba un índice que se editó en papel semestralmente hasta 1996 y se empleaba para búsquedas retrospectivas en papel. En estos índices puede apreciarse la complejidad de su clasificación jerárquica de tipo precoordinada.

En febrero de 2001 se puso en marcha la versión electrónica bajo la denominación de *elpais.es*, el cual junto con el resto de medios del grupo de comunicación *Prisacom*[10], utilizan un nuevo sistema editorial y documental basado en xml. Además, *Prisacom* optó por el estándar *News Industry Text Format* (*NIFT*) desarrollado por el *International Press Telecommunications Council* (*IPTC*) y el *Subject Reference System*. Esta clasificación temática se amplió y adaptó a las intereses de los medios de la empresa. **Flora Sanz** (2003) comenta que al tratarse de una clasificación muy general y con categorías propias del ámbito norteamericano se vieron obligados a incluir nuevos términos y añadir un tercer nivel de jerarquía para adaptarlo a sus necesidades. Por otra parte, se redujeron drásticamente las entradas utilizadas en la clasificación respecto a las empleadas en la base de datos del diario *El país* en versión papel. En el proceso de migración de los contenidos de las bases de datos en papel (*Hércules*) a la digital (*Pegaso*), se establecieron equivalencias entre las 750.000 entradas normalizadas de grandes temas, materias, topónimos, personas y empresas de *Hércule*s, y las 2.500 categorías[11] de *Pegaso*, basadas en la clasificación del *IPTC*.

"Sólo 2 de 20 bases de datos periodísticas online españolas analizadas por *Pastor* emplean un auténtico tesauro"

De los 11 servicios de documentación de medios de comunicación en España analizados por **Fuentes** y **Conesa** (1994) sólo en dos casos, *El observador* y el de la base *Basinfa* de *RTVE*, se comenta la utilización de tesauro. El diario catalán *El observador*, que dejó de publicarse en 1993, organizaba el archivo manual de prensa de acuerdo con un tesauro que se iba ampliando según las necesidades.

De entre las ponencias presentadas en el *Segon seminari: la documentació als mitjans de comunicació*[12], cerca de 20 trataban el caso concreto de algún medio de comunicación español (prensa, revistas, agencias, radio), y sólo en el ya citado caso de diario *El correo español* se hace referencia al empleo de lenguaje controlado. En las descripciones de las bases de datos *Baratz*, la revista *El Temps* y el archivo de ilustraciones del diario *ABC*, se refleja el empleo de un campo para descriptores, así como otro para los ámbitos geográficos si bien no se menciona el empleo de ningún tesauro. En relación a la base de datos de información de actualidad *Baratz*, **Aquesolo** (1995) señala que cada documento es analizado en un registro independiente al que se le asignan descriptores temáticos, de personalidades y geográficos. El mismo autor comenta otras aplicaciones de consulta de prensa y agencias, además de mencionar *Egunez egun* (*Día a día*), base de datos de prensa histórica del siglo XIX que fue desarrollada sobre el programa *Knosys* y que utiliza el tesauro *Eurovoc*.

La situación de los centros de documentación de prensa diaria en Andalucía ha sido estudiada por **Aquesolo** (1996). En su artículo expone que el sistema de clasificación de los fondos es básico: carpetas clasificadas por temas o según la estructura temática definida por las secciones del diario.

Sólo dos de veinte bases de datos periodísticas online españolas analizadas por **Pastor** (1992a) emplean estrategias de interrogación valiéndose de un auténtico tesauro. Se trata de la herramienta confeccionada por la empresa catalana *Enfony* y de la del Gobierno Vasco, *Inforpo 2*. *Enfony* ha confeccionado para cada base de datos diseñada a medida[13], un tesauro con los términos específicos de la temática que cubren. *Inforpo 2* usa también un tesauro referido a todos los temas de prensa recogidos y cuenta ya con más de 1.500 descriptores.

Por otra parte, las ediciones en cd-rom de los diarios *El mundo*, *La vanguardia* y *El periódico de Cataluña* tampoco disponen de tesauro. En cuanto al caso de *El país* añade una clasificación temática donde se organizan los conceptos, de más generales a más específicos. Tampoco las hemerotecas electrónicas de la mayor parte de los diarios

españoles de información general incorporan ningún tipo de vocabulario controlado para la recuperación de la información por parte de los usuarios. En el caso de la versión digital de *El país* tampoco se emplea ni en la hemeroteca ni en el buscador, pero sin embargo se puede apreciar la utilización de una clasificación jerárquica a partir del servicio denominado "El índice"[14], que incluye onomástico, geográfico, categorías temáticas y otras opciones.

En el entorno de la administración, **Izquierdo-Arroyo** y **Moreno-Fernández** (1992) describieron la estructura del tesauro diseñado para la base de datos de información de actualidad de *La región de Murcia*. Engloba 21 grandes categorías temáticas que incluyen la de onomásticos y topónimos. Se compone de dos partes: sistemática, incluyendo relaciones jerárquicas, partitivas, asociativas y de monoequivalencia semántica, así como notas de aplicación y sistema de facetas (agente, instrumento, modo, acción, materia) y otra alfabética.

Álvaro, **Villagrá** y **Sorli** (1989a) evaluaron 47 tesauros disponibles en lengua española. Para conseguir la relación de los mismos examinaron trece directorios y bibliografías sobre el tema. Ninguno de estos tesauros pertenece a un medio de comunicación.

Lo que si es frecuente es que algunos medios de comunicación españoles adapten tesauros ya elaborados, como el de la *Unesco* y el *Eurovoc*. El primero fue desarrollado por dicha organización para la indización y recuperación de información de la red integrada de documentación de este organismo. Fue publicado por primera vez en 1977 y en 1995 vio la luz una segunda edición. Está formado por unos 8.500 términos distribuidos en cinco secciones correspondientes a las principales áreas de actividad de la *Unesco*: educación, información y comunicación, ciencias sociales, cultura y humanidades, ciencia y tecnología y otra sección general. En ediciones posteriores se añadieron otras áreas de conocimiento: política, derecho y economía. Además, incluye nombres de países.

El *Eurovoc* es el tesauro empleado para la confección de los índices del *Diario oficial de las Comunidades Europeas*. Se caracteriza por su condición multidisciplinar, ya que cubre los temas de interés, tanto en el ámbito de la actividad parlamentaria, como en el de las instituciones comunitarias y en el de los estados miembros. Además añade dos secciones que contienen listas de topónimos y nombres de organizaciones.

Condicionantes para la utilización de tesauros en medios de comunicación

La clasificación e indización de los artículos periodísticos presenta unas características específicas que hacen difícil la aplicación de las clasificaciones alfabéticas de materias que se utilizan habitualmente en las bibliotecas generales, ya que se adaptan mal a la información periodística y la mayoría de los medios que elaboran dossiers o índices han desarrollado las suyas propias (**Fuentes**; **Conesa**, 1994). En este sentido **Díaz** et al. (1986) consideran que el problema de los tesauros existentes en ámbitos no periodísticos reside en ser sistemas destinados a la indización del conocimiento científico. En cambio, en la información de actualidad la materia prima son acontecimientos o reflexiones sobre los mismos.

García-Gutiérrez (1999) señala, refiriéndose a *la Clasificación Decimal Universal* y a lenguajes documentales similares: "si la CDU tuvo y sigue teniendo una gran aceptación, en el mundo biblioteconómico, para el control bibliográfico superficial del ámbito científico, la extrapolación de su filosofía a la organización documental del discurso periodístico sería un error ya que el enciclopedismo aparece como único rasgo común y tan sólo en el nivel extensional. De hecho el enciclopedismo que interesa al mass media queda marcado por intereses e ideología institucionales de los que la CDU carece a pesar de ser un producto del pensamiento positivista. Las restantes características de la actualidad eliminan la posibilidad de adoptar esquemas encorsetados, codificados, y de imposible puesta al día" (p. 356).

Pastor (1992) estima que las dificultades existentes en la indización de prensa escrita se derivan en su mayor parte de las características de la información de actualidad.

Rodríguez-Vela (1992) afirma que los tesauros de información de actualidad escasean dada la universalidad que se exige a su cobertura. En el mismo sentido **García-Gutiérrez** (1999) señala que la extensión enciclopédica de la actualidad crea dificultades a la hora de compilar y estructurar el vocabulario, aunque este problema queda compensado por la superficialidad de su tratamiento.

Además, la prensa nacional y regional adapta sus contenidos al ámbito geográfico al que pertenece. Este factor de localismo, que hace relevantes determinadas noticias en un contexto concreto, dificulta en cierta medida el compartir lenguajes controlados entre medios de comunicación distintos.

Perpinyà (1995) expone las condiciones específicas de las empresas del área de los medios de comunicación que condicionan las características principales de uso de los lenguajes documentales:

– El sector está constituido por empresas privadas y no centros de investigación, en consecuencia no existe una investigación conjunta

para la elaboración y utilización de un lenguaje común. Cada medio utiliza el lenguaje documental que le parece más adecuado.

– Los intereses en cada caso son muy distintos. Los documentos que se recogen y organizan en agencias de publicidad, prensa de información general, prensa deportiva, económica, del corazón, medios audiovisuales, etc, son absolutamente diferentes entre sí y requieren de instrumentos de descripción particulares.

– El tipo de información que procesan, de carácter general (cuestiones muy diversas) y de actualidad (se producen nuevos temas constantemente y otros se quedan obsoletos con rapidez) es otro condicionante esencial, puesto que obliga a tener un lenguaje muy flexible que admita muchos cambios.

– El usuario (básicamente el periodista) necesita obtener la información de forma inmediata. Esto influye en el tiempo dedicado al análisis documental, que tiene que ser mínimo, puesto que la base de datos se tiene que mantener permanentemente actualizada (p. 130).

A estas condiciones se puede añadir que la estructuración del sector en grupos de comunicación, de los que dependen conjuntos de cabeceras periodísticas o emisoras de radio y televisión, no se ha traducido en la existencia de centros de documentación que sirvan al conjunto de medios ligados empresarialmente. Tan sólo el empleo de las bases de datos de *El país* las empresas del grupo *PRISA* se puede citar como modelo de funcionamiento distribuido.

También **Perpinyà** (1995) comenta que la realidad de los sistemas de documentación periodísticos es que cada medio utiliza su propio lenguaje documental, y que generalmente se prefiere la indización a través del lenguaje libre o de listas de materia, debido al ahorro de tiempo en la construcción y utilización. Esto es así en las bases de datos de distribución pública, tanto de texto íntegro (*The guardian*, *The times*, *The sunday times*, *The independent*) como referenciales (*Baratz*, *Documentación de Medios*).

"Hay que plantear un acuerdo entre los medios de comunicación españoles para crear un vocabulario controlado común"

Otra de las razones para rechazar la construcción de un tesauro de información de actualidad es la exigencia de dedicación y tiempo que requiere su elaboración y su mantenimiento. **García-Gutiérrez** (1999) señala que la información y el vocabulario periodístico quedan obsoletos rápidamente por lo que cualquier lenguaje documental necesitaría de un equipo humano que realizara las actualizaciones constantemente y esto supone una asignación presupuestaria que no todos los medios de comunicación pueden permitirse.

En las conclusiones de un estudio sobre la situación de los centros de documentación de medios de comunicación de Madrid se señalan, entre otras, que la inexistencia de tesauros o lenguajes documentales especializados en información de actualidad ha originado que cada centro desarrolle su sistema propio dependiendo de necesidades y recursos muy concretos (**Razquin**, 1993).

Fuentes (1994) concluye, tras quince años de estudio de la documentación periodística, que no hay ninguna uniformidad ni intento de normalización de las operaciones documentales (selección, tratamiento y recuperación) que permita un intercambio de información entre los diferentes medios. Sin embargo, la iniciativa de un formato normalizado para textos en la industria periodística –*NITF* (*News Industry Text Format*) auspiciado por la *Newspaper Association of America* y el *International Press Telecommunications Council*)– no sólo incluye un formato normalizado para el intercambio de noticias, sino una clasificación temática recomendada, la *Subject Reference System.*

Por último, los estudios de consumo de información periodística revelan que buena parte de las necesidades expresadas se refieren a protagonistas de la actualidad y personajes públicos; por tanto, se basan en identificadores. **Castillo-Blasco** (2001) cuantifica en su estudio un 30,66% de peticiones basadas en identificadores, un resultado muy próximo al de **Iturregui** (comunicación personal, servicio de documentación escrita de *Euskal Telebista*).

Conclusiones y propuestas

La escasez de instrumentos de control terminológico en el tratamiento documental de la información de actualidad descansa en una serie de argumentos de naturaleza operativa de cierta solidez. Dichas cuestiones se basan en el elevado coste de la elaboración de tesauros, en la naturaleza dinámica y amplia cobertura temática del lenguaje periodístico, en la celeridad exigida a un espacio informativo de frecuencia tan elevada y en la naturaleza de las demandas planteadas, basadas en el "quién" y el "dónde" de las noticias.

Sin embargo, el ingente volumen de la información de actualidad y su creciente presencia en el entorno de la información accesible a través de internet, al alcance de usuarios no especializados, son argumentos que también aconsejan el control de su contenido. Además, la propia estructuración del trabajo periodístico,

con profesionales encuadrados en secciones o unidades especializadas, aconsejan el empleo de algún instrumento que permita organizar no sólo los documentos, sino también las peticiones. De esta forma se permitiría, por ejemplo, la implantación de sistemas de alerta destinados a los profesionales de los medios y también al público en general.

Desde este punto de vista, no parece descabellado plantear la necesidad de un acuerdo de mínimos entre los medios de comunicación españoles que se tradujera en un vocabulario controlado común para la organización de sus contenidos. Planteado en términos de macrotesauro, cada uno de los medios o grupos de comunicación tendría la posibilidad de extender ese instrumento en función de sus propias necesidades. Tomar como punto de partida los *iptc-subjectcode* de los *newscodes* del *Internacional Press Telecommunications Council* parece un arranque conveniente para abordar un esquema mínimo consensuado de control del lenguaje periodístico.

Agradecimientos:

A **Jesús Andérez** y a **Marta Iturregi** (*Dpto. de documentación y archivo*, *Euskal Telebista*) por los datos proporcionados sobre el archivo de prensa de su institución y la revisión crítica del manuscrito.

Notas

1. Cifras del *Anuario El país* 2005.

2. El *Subject Reference System* está incluido en la *IIMG*: *Information Interchange Model Guideline*. Tiene 3 niveles de jerarquía.

3. *http://www.ina.fr/CP/MondeDiplo/Thesaurus/thesaurus.fr.htm*

4. *http://www.darmstadt.gmd.de/~lutes/thesoecd.html*

5. *http://www.iptc.org/IIM/3.0/specification/IIMV3.PDF*

6. Actualmente *AIRS*.

7. *http://www.fiatifta.org*

8. El *Grupo Correo* se fusionó en 2001 con el *Grupo Prensa Española* y a partir de 2003 pasó a denominarse *Vocento.*

9. Editor de los diarios *La vanguardia*, *Mundo deportivo* y revistas como *Magazine*, *Què fem?*, *Què más?* y el semanario del motor *Escape*.

10. *El país*, *As*, *Cinco días*, *Cadena SER* y *Los 40.com*.

11. El total de términos que incluye esta segunda clasificación es de aproximadamente 10.000 si se tienen en cuenta también los términos geográficos y los onomásticos de personas y empresas.

12. Celebrado en Valencia del 7 al 9 de marzo de 1994.

13. Estas bases de datos recopilan una selección de artículos de prensa sobre un tema concreto de interés para el cliente.

14. Con esta opción se recupera a partir de los índices las noticias de ese día que coinciden con el epígrafe del índice solicitado. En el caso de los onomásticos y geográficos se permite ampliar la búsqueda a otros documentos del retrospectivos del archivo. En todos los casos la visualización del documento final es solo para suscriptores.

Referencias bibliográficas

Anuario El país 2005. Madrid: Ediciones El País, 2005. ISBN 84-95595-12-5.

Aguado-González, Francisco-Javier. "Organización del sistema de archivo y documentación de Prensa Española, S. A. (ABC y Blanco y negro)". En: *Revista general de información y documentación*, 1995, n. 2, pp. 203-208.

Álvaro-Bermejo, Concha; **Villagrá-Rubio, Ángel**; **Sorli-Rojo, Ángela**. "Desarrollo de lenguajes documentales formalizados en lengua española: evaluación de los tesauros disponibles en lengua española". En: *Revista española de documentación científica*, 1989, n. 3, pp. 283-305.

Aquesolo-Vegas, José. "Situación de los servicios de documentación de la prensa diaria de Andalucía". En: *Cuadernos de documentación multimedia,* 1996, n. 5. Consultado en: 20-04-00. *http://www.ucm.es/info/multidoc/multidoc/revista/cuadern5/aquesolo.htm*

Calvanese, Diego; **Catarti, Tiziana**; **Santucci, Giuseppe**. "Laurin: a distributed digital library of newspaper clippings". En: *World wide web*, 2001, n. 4, pp. 5-20.

Caridad-Sebastián, Mercedes. "Estructura general del banco de datos del New York times". En: *Documentación de las ciencias de la información*, 1980, n. 4, pp. 139-155.

Castillo-Blasco, Lourdes; **Doménech-Vidal, Soledad**; **Soler-Monreal, Concepción**; **Amat, Carlos B.** "Demanda de información de actualidad en un servicio de referencia periodística. Análisis descriptivo de 4.160 solicitudes". En: *Revista española de documentación científica*, 2001, v. 24, n. 1, pp. 36-50.

Codina, Lluís. *Teoria de sistemes, teoria de recuperació d'informació i documentació periodística*. Barcelona: Universitat Autònoma de Barcelona, 1996. ISBN 84-490-0725-9; 978-84-490-0725-5.

Coll-Vinent, Roberto. *Teoría y práctica de la documentación*. Barcelona: A.T.E., 1978. ISBN 84-7442-030-X; 978-84-7442-030-2.

Coll-Vinent, Roberto. *Teoría de la teledocumentación*. Barcelona: A.T.E., 1982. ISBN 84-7442-164-0; 978-84-7442-164-4.

Díaz-Arias, Rafael, et al. "La base de información de actualidad (Basinfa) de RTVE, un sistema automatizado de documentación periodística". En: *Segundas jornadas españolas de documentación automatizada*, 1986, pp. 175-183.

Díaz-Nosty, Bernardo. **"**La difusión de la prensa diaria en lengua española". En: *El español en el mundo. Anuario del Instituto Cervantes 1999*. Barcelona: Plaza y Janés, pp. 65-130. Consultado en: 04-11-06. *http://cvc.cervantes.es/obref/anuario/anuario_99/*

Fuentes-Pujol, Maria-Eulàlia. "Evolució de la documentació periodística a Espanya durant els darrers cinc anys i algunes experiències europees". En: *Segon seminari: la documentació als mitjans d'informació. L'experiència multimèdia. Ponències i conclusions*, 1994, pp. 17-28.

Fuentes-Pujol, Maria-Eulàlia; **Conesa, Alicia**. *La documentació periodística. Catalunya, Espanya i altres experiències europees*. Barcelona: Generalitat de Catalunya, 1994.

Galdón-López, Gabriel. *Perfil histórico de la documentación en la prensa de información general 1845-1984.* Pamplona: Eunsa, 1994.

Galdón-López, Gabriel. *El servicio de documentación de prensa: funciones y métodos.* Barcelona: Mitre, 1986.

García-Gutiérrez, Antonio-Luis. "Lenguajes documentales e información de actualidad". En: **García-Gutiérrez, A. L.** (ed.). *Introducción a la documentación informativa y periodística*. Sevilla: Editorial MAD, 1999, pp. 351-372.

García-Morales, Justo. "Normas para la preparación de índices analíticos en centros de documentación y archivos de publicaciones periódicas". En: *Documentación de las ciencias de la información*, 1979, n. 3, pp. 71-111.

Gómez, Bernardo; **Paniagua, Francisco**. "Las ediciones digitales de los diarios españoles. Nacimiento y consolidación de un sector en auge". En: *Razón y palabra*, n. 47, 2005. Consultado en: 04-11-06. *http://www.cem.itesm.mx/dacs/publicaciones/logos/anteriores/n47/gomezpaniagua.html*

González-Quesada, Alfons. "La evolución histórica de la documentación periodística". En: **Fuentes-Pujol, M. E.** (ed.). *Manual de documentación periodística.* Madrid: Síntesis, 1995, pp. 23-39.

Greengrass, Alan R. "Indexing at the New York times information service". En: *Indexing specialized formats and subjects*. Metuchen: Scarecrow Press, 1983, pp. 180-188.

Guallar, Javier. **"**iMente, servicios de información de actualidad en línea". En: *El profesional de la información*, 2006, v. 15, n. 6, pp. 426-435. Consultado en: 04-01-07. *http://eprints.rclis.org/archive/00007856/01/epi06_guallar_imente.pdf*

Guallar, Javier. **"**Mètodes i tècniques de recerca en els articles de documentació periodística a Espanya (1997-2002)". En: *BiD*, 2003, n. 11. Consultado en: 04-01-07. *http://www2.ub.es/bid/consulta_articulos.php?fichero=11gualla.htm*

Hanse, Kathleen A.; **Ward, Jean**. "Information technology changes in large newspaper libraries". En: *Special libraries*, 1991, v. 82, n. 4, pp. 267-273.

Hegg, Judith L. "Small newspaper libraries: the libraries that time (and automation) passed by". En: *Special libraries*, 1991, v. 82, n. 4, pp. 274-281.

International Press Telecommunications Council. *Information Interchange Model Guideline 3*, 1999. Consultado en: 19-07-01.
http://www.iptc.org

International Press Telecommunications Council. *Subject Reference System Guidelines. Versión 3*, 2003. Consultado en: 21-02-04.
http://www.iptc.org

Izquierdo-Arroyo, José-María; **Moreno-Fernández, Luis-Miguel**. "Diseño de una base de datos de prensa controlada por un lenguaje facetado de estructura combinatoria ("thesaurus")". En: *Revista española de documentación científica*, 1992, v. 15, n. 1, pp. 44-63.

Jiménez, Àngels; **González, Alfons**; **Fuentes-Pujol, Maria-Eulàlia**. "Las hemerotecas digitales de la prensa en internet". En: *El profesional de la información*, 2000, v. 9, n. 5, pp. 15-24.

Llobet, Montserrat; **Pañella, Imma**. "Un thesaurus aplicat a un mitjà de comunicació audiovisual. Experiència a TV3". En: *Item*, 1988, n. 2-3, pp. 51-60.

Martín, Mauricio. "El correo. Documentación de prensa. Aprovechando recursos". En: *Segon seminari la documentació als mitjans d'informació. L'experiència multimèdia. Ponències i conclusions*, 1994, pp. 54-62.

Martín-Muñoz, Javier; **López-Pavillard, Jacobo**. "La documentación audiovisual en RTVE". En: *Documentación de las ciencias de la información*, 1995, n. 18, p. 143-171.

Martínez-Pestaña, M. Jesús. "Estructura de los bancos y bases de datos de prensa". En: *Documentación de las ciencias de la información*, 1986, n. 10, pp. 159-212.

Milstead, Jessica L. "Newspaper indexing: The Official Washington Post Index". En: *Indexing specialized formats and subjects*. Metuchen: Scarecrow Press, 1983, pp. 189-204.

Mühlberger, Günter. "Newspaper clippings in a digital world: the Laurin project". En: *Exploit interactive*, 1999, n. 2, 20 July. Consultado en: 13-04-00.
http://www.exploit-lib.org/issue2/laurin/

Pastor-Ruiz, Fátima. *La irrupción de las nuevas tecnologías en la documentación periodística*. Bilbao: Universidad del País Vasco, 1992.

Perpinyà-Morera, Remei. "Los lenguajes documentales". En: **Fuentes-Pujol, M. E.** (ed.). *Manual de documentación periodística*. Madrid: Síntesis, 1995, pp. 111-132.

Razquin, Pedro. "Situación de los centros de documentación en los medios de comunicación de Madrid". En: *Cuadernos de documentación multimedia*, 1993, n. 2. Consultado en: 30-06-00.
http://www.ucm.es/info/multidoc/multidoc/revista/num2/prazquin.html

Rodríguez-Vela, Cristina. "Los lenguajes documentales en las bases de información política y de actualidad". En: *Revista española de documentación científica*, 1992, n. 1, pp. 13-23.

Sanz-Calama, Flora. "La hemeroteca digital de El país". En: *IV jornadas de bibliotecas digitales*, 2003. Consultado en: 21-06-04.
http://imhotep.unizar.es/jbidi/jbidi2003/14_2003.pdf

Semonche, Barbara P. "Newspaper indexing policies and procedures". En: *News media libraries. A management handbook*. Westport: Greenwood Publishing Group, 1993. Consultado en: 19-01-02.
http://metalab.unc.edu/journalism/indexing.html

Whatmore, Geoffrey. *La documentación de la noticia: organización y métodos de trabajo para archivos de referencias de periódicos y agencias.* Pamplona: Universidad de Navarra, 1990.

Wellisch, Hans H. *Indexing from A to Z.* New York: Wilson Company, 1991.

Lourdes Castillo, *Unidad de Documentación de RTVV, Pista de Ademuz s/n, 46100 Burjassot, Valencia.*
macas@uv.es

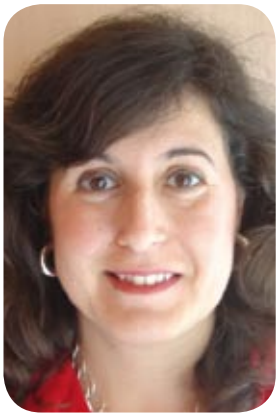
Alejandro De-la-Cueva*, Departamento de Historia de la Ciencia y Documentación, Facultad de Medicina y Odontología, Av. Blasco Ibáñez 15, 46010 Valencia.*
alejandro.cueva@uv.es

Aplicación de un nuevo sistema de indización en una colección de recursos especializados en ciencias de la educación

Por **Mariàngels Granados** y **Anna Nicolau**

Resumen: *Este estudio sobre una colección de ocho documentos sobre enseñanza primaria (ciencias de la educación) sirve de marco para presentar un nuevo sistema de indización, basado en una estructura de descriptores primarios y secundarios y su equivalencia con la Clasificación Decimal Universal. Se trata de poner de relieve las ventajas de su uso en dos vertientes: desde la perspectiva del sistema de búsqueda de la información y su implicación en el sistema y proceso de indización. Las directrices que rigen en este caso práctico explican el sistema de indización postcoordinado empleado en base a la técnica de ordenación nuclear de los descriptores. Para la gestión de la información en el contexto marcado, se plantea la necesidad de un sistema informático, del que también se indican una serie de requisitos que faciliten la explotación de sus posibilidades deconsulta y recuperación de información relevante.*

Palabras clave: *Lenguajes de indización postcoordinados, Clasificación Decimal Universal, Indización, Enseñanza primaria, Ciencias de la educación, Tesauros, Sistemas de clasificación, Catálogos en línea, Gestión de la información, Sistemas de recuperación de la información.*



Mariàngels Granados *cursa la licenciatura en documentación en la Universitat de Barcelona. En su experiencia profesional destacan: Biblioteca del Col·legi d'Odontòlegs i Estomatòlegs de Catalunya, Centre de Documentació Juvenil de la Generalitat de Catalunya (Archivo, Sección de Prensa, Biblioteca i Exposición Permanente), Hemeroteca Nacional de Catalunya,y desde 1994, Biblioteca de Catalunya (Unitat Bibliogràfica, Col·leccions Generals).*

Anna Nicolau *es diplomada en biblioteconomía y documentación y cursa la licenciatura en documentación por la Universitat de Barcelona. Centra su trayectoria profesional en la organización, clasificación y catalogación de distintas tipologías documentales y la experiencia en diferentes centros de la administración pública, entidades privadas y universidades.Trabaja en la Biblioteca de Catalunya desde 1999.*

Title: Implementation of a new indexing system for a specialized resource collection in teacher preparation

Abstract: *A collection of eight documents on primary education provides the example for presenting a new indexing system based on a structure of primary and secondary descriptors and illustrating its equivalence to Universal Decimal Classification. The article shows the advantages of this new system as an information retrieval system and describes its role in the indexing system and process. Another aim of this article is to describe an actual project using a postcoordinate indexing system based on nuclear ordering of the descriptors, as well as to define the system and software requirements for efficient operation of user search and information retrieval functions.*

Keywords: *Postcoordinate indexing languages, Thesauri, Classification systems, Universal Decimal Classification, Content analysis, Web catalogs, Information management, Information retrieval systems, Education.*

Granados, Mariàngels; Nicolau, Anna. *"Aplicación de un nuevo sistema de indización en una colección de recursos especializados en ciencias de la educación". En: El profesional de la información,* 2007, noviembre-diciembre, v. 16, n. 6, pp. 627-635.

DOI: 10.3145/epi.2007.nov.09

Introducción

ESTE TRABAJO CONSTA DE DOS PARTES: En primer lugar hemos situado el contexto de la colección de referencia que se ha indizado y resumido, especializada en ciencias de la educación y con diferentes tipos documentales. Hemos seleccionado este tema por nuestro conocimiento de la disciplina a nivel de indización y clasificación.

En esta primera parte se ha seguido el siguiente método: la descripción del sistema de indización piloto que aplicaríamos a los descriptores y el análisis de contenido y selección de los términos de algunos documentos para comprobar su funcionamiento. Así hemos constatado que el método era idóneo y hemos continuado explicando los tesauros utilizados. Se estipulan unas concreciones sobre el uso y orden

de los elementos y del sistema de clasificación aplicado (*CDU*).

En la segunda parte del artículo se tratan las características que precisa el sistema informático de gestión documental para hacer posible el tratamiento y la recuperación de la información de acuerdo con las premisas establecidas.

1. Contexto

El perfil de centro del caso práctico es uno de carácter especializado en ciencias de la educación del grado educativo de primaria, secundaria y bachillerato en España, cuyos objetivos serían: la investigación educativa, la psicología y sociología de la educación, la formación del profesorado, la dinamización pedagógica, el apoyo a la docencia y las actividades preescolares. La colección incluye documentos de diferentes tipos según criterios tales como sus características físicas, el contenido o su naturaleza (original, copias impresas o seriado). La audiencia prevista en una unidad de información de tales características tiene en cuenta el profesorado, investigadores y profesionales de áreas de especialidad afines como pedagogos, psicólogos, trabajadores sociales y educadores.

Si bien el centro en base al cual hemos efectuado las indizaciones es especializado, tenemos que remarcar el ámbito multidisciplinar en el que ubicamos la recuperación de la información según el sistema expuesto.

2. Sistema de indización

2.1. Análisis

En el proceso de indización la coherencia es un factor clave, puesto que la subjetividad condiciona la selección de descriptores por parte de un mismo profesional en distintos momentos y más incluso cuando se trata de la misma catalogación realizada por diferentes analistas. Aunque se pretenda alcanzar una buena política de indización, el nivel de coherencia nunca suele ser tan elevado como nos gustaría. En EUA, este porcentaje se sitúa en el 60% como máximo, aunque habitualmente es del 25%. Así, esta incoherencia deja de ser una anomalía y pasa a ser un hecho normal. Con frecuencia, el grado de coherencia es muy alto en los dos primeros términos seleccionados, motivo por el cual acostumbra a coincidir con los conceptos principales.

2.2. Sistema de indización

Nuestro planteamiento de nuevo sistema de indización está basado en la utilización de descriptores y de la *Clasificación Decimal Universal* (*CDU*). La *CDU* resulta ser el equivalente o la transliteración de un sistema de descriptores primarios asignados de una manera lógica y ordenada con el fin de representar el contenido principal del documento.

Las ventajas de este nuevo sistema quedan dibujadas por lo que denominamos "índices permutados" de descriptores y de la *CDU*. Estos índices son los más adecuados para evidenciar la sistematización de este proceso que relaciona los temas y números de las notaciones decimales, gracias a la frecuencia de uso en generarse, y que supone una garantía para solucionar algunos de los principales problemas de la indización en el mundo bibliotecario, como la coherencia, la selectividad o la especificidad y la uniformidad, conceptos que explicaremos más adelante.

Además, los índices facilitan la comparación de los términos con las signaturas correspondientes a un mismo concepto por su proximidad, y a la vez representan la frecuencia en la que han sido asignados y combinados por parte de los profesionales durante el proceso de catalogación. Esta frecuencia permite también traslucir la correspondencia (que en un índice real se hace patente de forma más clara) entre las materias y los códigos numéricos elaborados con la clasificación decimal. Así pues, se puede saber si un conjunto de descriptores se ha utilizado junto con una misma notación para representar idéntico concepto. Y es este hecho el que proponemos marcar como pauta a seguir (en relación a los principios de procedimiento que antes hemos mencionado), dado que siempre se procurará actuar del mismo modo. En este caso podemos demostrar la potencia de este sistema de indización para conseguir la máxima coherencia posible. La uniformidad del catálogo queda entonces garantizada puesto que el contenido de los documentos que tratan de los mismos conceptos se indizará con los mismos términos controlados.

"El proceso descrito permitirá controlar y ordenar todos los conceptos que expresen un determinado contenido de manera lógica y será siempre exhaustivo"

Mediante la correspondencia del conjunto de descriptores con la *CDU* también conseguimos un alto grado de especificidad. No habrá ningún conjunto de descriptores primarios que representen un nivel más general ni más específico que el que represente la notación asignada. El proceso descrito permitirá controlar y ordenar todos los conceptos que expresen un determinado contenido de manera lógica y será siempre exhaustivo. No obstante, la política del centro dónde se aplique deberá marcar su uso en mayor o menor grado en función de la colección, necesidades y objetivos. La correlación en línea de materias y notaciones de clasificación podría ayudar a normalizar y conseguir la uniformidad esperada.

En relación a la recuperación, la búsqueda por materia o palabra clave nos llevaría a identificar su número correspondiente, como siguiente paso. La equivalencia mencionada se obtendría de la permutación de los índices de encabezamientos de materia a *CDU*, y también de *CDU* a encabezamientos de materia recíprocamente. Estos índices ponen de manifiesto las interacciones de los lenguajes documentales empleados y permiten ver la frecuencia de uso de la notación de todos los documentos de la base de datos.

El procedimiento en cadena ideado por **Ranganathan**[1] ha sido muy estudiado y desarrollado en la actualidad: una de las especialistas es **Rosa San-Segundo** de la *Universidad Carlos III de Madrid,* que tiene extensa bibliografía publicada sobre la indización y la clasificación en cadena. Nuestro estudio se fundamenta en esta técnica, pero además de coherencia pretende conseguir un nivel de relevancia más alto mediante la ponderación de los descriptores al separar en dos los campos de descriptores primarios y los secundarios (los tratados con menor intensidad o en parte del documento) y darles un orden lógico.

2.3. Tesauro

En la colección de referencia especializada en ciencias de la educación, hemos optado por una combinación de lenguajes postcoordinados: un tesauro que elaboraríamos adaptando los ya existentes en este campo a la realidad concreta:

– *ERIC (Educational Resources Information Center): processing and reference facility*, 2003.

http://www.eric.ed.gov/

– **Houston, James E**. *Thesaurus of ERIC descriptors*. 12th ed. Phoenix: Oryx, 1990.

– *Comissió de les Comunitats Europees, i Consell d'Europa*. *Tesaurus europeu de l'educació*. Versió en llengua catalana, 2003.

http://www.mec.es/redinet2/html/TEECAT.pdf

ERIC es un sistema de información patrocinado por el *Departamento de Educación* de los EUA y el *Institute of Education Science*. En nuestro estudio seguiremos este tesauro como fuente de referencia (previa traducción y adaptación al contexto). Hemos consultado la versión en línea de 2003 en combinación con la impresa del año 1995. La base de datos que utiliza este lenguaje estructura la información de la forma que puede verse en la figura 1.

Los descriptores precedidos por un asterisco son primarios ("Major descriptors") y el resto son secundarios ("Minor descriptors") (*ERIC*, 1995). Los identificadores, como veremos, se ofrecen en otro campo al igual que el tipo de documento cuando no es materia. En cuanto a los descriptores observamos que están todos mezclados sin ningún orden aparente. Nuestra base de datos, en cambio, dispondría la información como puede verse en la figura 2.

Según el ejemplo, éstos serían los términos de indización correspondientes a un documento que nos hable principalmente de propuestas didácticas y de actualización científica sobre aritmética para maestros y ofrece como contexto la reforma educativa en Catalunya haciendo referencia sólo superficialmente a la italiana. Como veremos, el sistema informático que emplearemos también tendrá que permitir realizar la búsqueda según su relevancia. Si tenemos en cuenta la visualización de la descripción del documento, los diferentes elementos siguen un orden y el registro reúne en un mismo campo descriptores e identificadores.

Esta disposición no responde al orden en que aparecen en el documento, sino al orden "nuclear" de los conceptos principales o más relevantes que expresan los documentos. Para los descriptores secundarios habrá otro campo que se creará siguiendo el criterio expuesto.

Se ha considerado por orden de relevancia indicar los identificadores al final de la secuencia, concretamente los geográficos y cronológicos. Sin embargo, esto no impide que todos los términos queden indizados separadamente y ofrezcan el mismo potencial de búsqueda:

– Descriptores extraídos de los lenguajes controlados citados anteriormente.

High School Foreign Language Programs: A renewed Challenge.
Descriptors: Secondary Education; *Second Language Instruction; *Cultural Awareness; Cultural Education; Critical Thinking; Worksheets; Cross Cultural Studies; Reading Instruction; German
Identifiers: Cultural Literacy
Pub Type:

Figura 1

Aritmética 2
Descriptores primarios:* Aritmética ; Enseñanza primaria ; Innovación curricular
Descriptores primarios:* Aritmética ; Enseñanza primaria ; Didáctica
Descriptores secundarios:Reforma educativa ; Catalunya ; 1990
Descriptores secundarios:Reforma educativa ; Italia ; 1990
Tip. *Pub:* Video
CDU: 511:371.014.5 ; 511:373.43

Figura 2

37.014.6(467.1 Mr Alella)	Escuela pública; Evaluación; Alella (Catalunya); 1998
37.057CEIP Fabra(467.1 Mr Alella)	Escuela pública; CEIP Fabra; Alella (Catalunya); 1998
371: 681.3(467.1)	Enseñanza; Nuevas tecnologías; Catalunya
371.014.1(467.1 Ba Badalona)"1971"	Enseñanza primaria; Igualdad de oportunidades; Badalona (Catalunya); 1971
371.014.53(467.1 Ba Badalona)"1971"	Enseñanza primaria; Desigualdad social; Badalona; 1971
371.12:681.3(467.1)	Profesión docente; Red telemática; Catalunya
371.263(467.1 Ba Badalona)"1971"	Enseñanza primaria; Evaluación inicial; Tests de diagnóstico; Badalona (Catalunya) ;1971
373.3/.47(467.1)	Enseñanza primaria; Catalunya
373.3(094.58)(460)"18"	Enseñanza primaria; Sistema educativo; Legislación; España; XIX
373.3.078(467.1)	Enseñanza primaria; Escuela pública; Administrador; Catalunya
511:371.014.5	Aritmética; Enseñanza primaria; Innovación curricular
511:373.43	Aritmética; Enseñanza primaria; Didáctica
78:371.3Mètode Willems i Chapui	Método Willems y Chapui; Enseñanza primaria
78:373.3.02	Educación musical; Enseñanza primaria
804.99:373.3(467.1 Ba Badalona)"1971"	Lengua catalana; Enseñanza primaria; Badalona (Catalunya); 1971

Tabla 1. Ejemplo de índice permutado de CDU/descriptores.
Los diferentes descriptores quedan ordenados según el orden decimal del 0 al 9.

– Identificadores para representar y designar nombres de personas y entidades, geográficos, períodos de tiempo y las tipologías documentales.

3. Sistema de clasificación

Para la clasificación documental se sigue la versión expandida de la *CDU*, en este caso la subclase 37:

– *Universal Decimal Classification (BSI Standards)*. English full ed. London: British Standards Institution, 1975.

Cuando haga falta optar por la relación de notaciones, se recurrirá a la edición abreviada española:

– *CDU: Clasificación Decimal Universal*. Madrid: Aenor, 1995.

Para subdivisiones geográficas de topónimos catalanes:

– **Rubió i Balaguer, Jordi**, dir. *Classificació decimal per a les biblioteques catalanes*. Barcelona: Teide, 1982.

Los argumentos que podemos aportar para justificar el importante papel de las clasificaciones en la recuperación de la información en sistemas automatizados vienen a ser los que se listan a continuación:

– Capacidad de combinar la búsqueda por palabra clave (recuperación) con la búsqueda por índice (navegación). Los sistemas de clasificación permiten la visualización por pantalla de los registros temáticamente afines. Además, el esquema puede ser utilizado como posible interfaz de cara a la consulta general de un catálogo por materias (por ejemplo, la *Biblioteca Virtual Cervantes*).

– Contextualización de las búsquedas cubiertas por un campo semántico. Este aspecto resulta especialmente práctico cuando se combina con la recuperación por palabras clave a la hora de reducir los problemas de ambigüedad del lenguaje natural.

– Modificar la estrategia de búsqueda, ampliándola o limitándola de un nivel más genérico a otro más específico gracias a la capacidad de navegación implícita de la organización jerárquica de la información, a partir del código resultante asignado a un registro bibliográfico conocido.

– Las notaciones pueden funcionar como un lenguaje puente para contrarrestar y sobrepasar barreras lingüísticas en bases de datos muy amplias y con registros en diferentes idiomas.

– La organización de los conceptos está basada en las relaciones de materia y no en el alfabeto. Esto hace que los temas dentro de una disciplina queden agrupados en lugar de quedar dispersos (utilizando un orden aleatorio o más bien convencional).

Las ventajas que se han valorado de la *CDU* son además:

– La simplificación en la repre-

Aritmética; Enseñanza primaria; Innovación curricular	511:371.014.5
Educación musical; Enseñanza primaria	8:373.3.02
Enseñanza; Nuevas tecnologías; Catalunya	371: 681.3(467.1)
Enseñanza primaria; Evaluación inicial; Tests de diagnóstico;	371.263(467.1Ba Badalona)"1971"
Badalona (Catalunya); 1971 Enseñanza primaria; Catalunya	373.3/.47(467.1)
Enseñanza primaria; Desigualdad social; Badalona; 1971	371.014.53(467.1Ba Badalona)"1971"
Enseñanza primaria; Escuela pública; Administrador; Catalunya	373.3.078(467.1)
Enseñanza primaria; Igualdad de oportunidades;	
Badalona	371.014.1(467.1Ba Badalona)"1971"
(Catalunya); 1971	
Enseñanza primaria; Sistema educativo; España; XIX	373.3(094.58)(460)"18"
Escuela pública; Evaluación; Alella (Catalunya); 1998	37.014.6(467.1 Mr Alella)
Escuela pública; CEIP Fabra; Alella (Catalunya); 1998	37.057CEIP Fabra(467.1 Mr Alella)
Lengua catalana; Enseñanza primaria;	
Badalona (Catalunya); 1971	804.99:373.3(467.1 Badalona)"1971"
Método Willems y Chapui; Enseñanza primaria	78:371.3Mètode Willems i Chapui
Professió docent; Red telemática; Catalunya	371.12:681.3(467.1)

Tabla 2. Ejemplo de índice permutado de descriptores / CDU
El juego de descriptores primarios de los documentos quedan ordenados alfabéticamente.

sentación de los conceptos, como aplicación de una sintaxis para relacionar los términos.

– La construcción de notaciones equiparables a la cadena de descriptores.

– La flexibilidad con respecto a la combinación de elementos y a la ordenación de algunas subdivisiones para adaptarse al orden nuclear y a las necesidades locales.

– La cobertura universal de todas las materias, de manera más o menos desarrollada.

– La organización y visualización de manera que permite destacar las relaciones entre las diversas áreas temáticas.

– La polijerarquía, valorada comouna característica positiva de su estructura, pues la asignación de más de un término genérico a cada descriptor contribuiría a dar flexibilidad al lenguaje documental para representar los diferentes enfoques o puntos de vista sobre los cuales puede ser tratado un mismo tema.

– La organización de algunos conceptos por facetas (subdivisiones especiales).

El principal inconveniente que apuntan los teóricos[2] y que hemos constatado en la utilización de la *CDU* es la necesidad por parte del usuario de saber el número concreto que pertenece a su concepto y, por lo tanto, la necesidad de conocer su estructura, su funcionamiento y el no disponer de una versión en línea actualizada y de acceso gratuito (*UDC Consortium*, en línea). Este problema se puede sortear si hacemos la primera consulta por descriptores mediante la estructura de navegación o bien por palabra clave ya que entonces los índices nos enviarán al número concreto de *CDU*, que a su vez nos servirá para ampliar la búsqueda a los conjuntos de descriptores primarios que sean equivalentes.

La opción complementaria consistiría en prescindir del índice temático de *CDU* y utilizar enlaces de dos niveles: unos serian encabezamientos temáticos equivalentes a las notaciones de *CDU*, establecidos como clases y los otros serian los descriptores que representan el contenido de los documentos. En estos casos, el encabezamiento correspondiente a las clases y subclases debería diferenciarse tipográficamente del resto. Su objetivo: contrarrestar la dificultades para los usuarios derivadas del uso de la *CDU* y contextualizar el usuario en el ámbito multidisciplinar correspondiente en cuanto al sistema de recuperación. De este modo, no sería imprescindible conocer los códigos y directamente el acceso se haría por medio de la información textual, por lo que entonces recurriríamos a las subdivisiones especiales de cada clase para una estructura facetada en la que se contemplarán

entre otras[3] los identificadores con connotación de tiempo, tipo documental, punto de vista y lugar.

Durante la elaboración del presente trabajo han surgido interrogantes sobre la visualización de los documentos con notaciones relacionadas. Este problema se ha sorteado con la previsión de la técnica de ubicación múltiple que hemos estipulado y creemos que debería tener el sistema de gestión documental. Otra cuestión planteada ha sido la gestión de las palabras clave con las que el usuario consulta el sistema. La decisión ha consistido en canalizarlas con los descriptores y encabezamientos afines consiguiendo así mediante estos enlaces (tanto de palabras clave como entre las clases establecidas) un amplio acceso a la información y facilitar la navegación, consideraciones que deberá contemplar el sistema informático.

4. Sistema informático de gestión documental

El sistema de recuperación de información (SRI) empleado debería ser integrado y escalable para que permitiera llevar a cabo todas las tareas de gestión bibliotecaria y de manera progresiva, por módulos; entre ellas: la catalogación y el control de autoridades, consulta (opac), adquisiciones, circulación (préstamo, reservas, prórrogas, etc.), control y seguimiento de las publicaciones periódicas, consulta del directorio de usuarios, edición y creación de estadísticas y de otros listados (por ejemplo, el de novedades o bibliografías para la difusión selectiva de la información, según perfiles de usuarios).

4.1. Módulo de catalogación y indización

A continuación se listan las posibilidades que ofrecerá y deberá garantizar el sistema:

– Admitir el formato *MARC21* para la descripción bibliográfica y visualización de resultados.

– Dejar exportar o importar registros en este mismo formato.

– Ejercer control de los ejemplares: numeración a partir de códigos de barras y código topográfico.

– Copiar registros con la finalidad de crear otros registros nuevos.

– Diseñar varias plantillas de entrada de datos según modelos por tipologías de documentos.

– Corregir automáticamente una autoridad en todos los registros bibliográficos asociados a la vez.

– Indizar aquellos conceptos que necesiten de más de un término para ser representados (cadenas de descriptores): por ejemplo, nombre-adjetivo con connotación diferente a la suma de los dos conceptos por separado.

– Ponderar los descriptores (descriptores primarios y descriptores secundarios).

– Hacer enlaces a documentos electrónicos/digitales relacionados: imágenes, ayudas de búsqueda, fuente, etc.

– Trabajar con diferentes catálogos que formen un todo (un catálogo formado por varios grupos o clases).

– Gestionar el tesauro para establecer referencias que relacionen los diferentes términos de indización y sin que generen incompatibilidades (referencias de véase, véase también y jerárquicas).

– Crear automáticamente índices de los campos: autor, título, materias (descriptores e identificadores), palabras clave y tipos de documentos. Aparte, poder definir el fichero de palabras vacías.

– Tener en cuenta los campos de: "Descriptores", "Identificadores", "CDU" y "Área temática" (notación/nes equivalente/s a los conceptos nucleares y que podrían ser más de uno en los casos de polijerarquía o doble dependencia jerárquica)[4].

– Hacer búsquedas postcoordinadas por los elementos de la *CDU* que conforman las notaciones, sin necesidad de introducirla desdoblada (por separado) en el caso de las que estén relacionadas por medio del símbolo específico (:).

– Introducir en el sistema los sumarios sólo de las obras colectivas entendiendo como tales las compilaciones y las publicaciones periódicas especializadas: revistas, boletines, anuarios, jornadas, congresos y seminarios.

– Gestionar las autoridades de *CDU*.

4.2. Módulo de consulta y recuperación

Requisitos:

– El módulo de administración debe permitir modificar los índices para añadir, eliminar o corregir, una vez implementado el sistema.

– Integrar el control automático de autoridades con las relaciones jerárquicas y juegos de referencias pertinentes.

– Los lenguajes de indización/recuperación deberán ser consultables para los usuarios de manera que puedan guiarlos y asistirlos durante el proceso de búsqueda por navegación: se requeriría el acceso al tesauro y a los índices permutados de materias y notaciones que definen el sistema.

– Poder visualizar todos los campos informativos empleados en las plantillas de entrada de datos, incluida la totalidad de las notas.

– Capacidad de navegación a partir de cada uno de los descriptores y las notaciones.

– Seleccionar los campos por los que se podrá consultar el catálogo: autor, título, materia, año, palabra clave, topográfico, etc.

– Implementar un motor de búsqueda interno en la base de da-

tos que permita recuperar a texto completo y de forma ponderada sólo en los campos de título, resumen, identificadores y sumarios. Y también contemplar la opción de búsqueda avanzada por campos tal cual, en general, haciendo distinción entre descriptores primarios y secundarios con el fin de: extraer listados o índices de palabras clave a partir de los datos internos; hacer búsquedas utilizando los operadores booleanos; permitir el truncamiento de las palabras; eliminar o más bien prevenir, en definitiva, el ruido documental en los resultados.

– Recuperación a partir de todos los términos específicos de cada término genérico al cual están estos subordinados.

– Posibilidad de ampliar o limitar la búsqueda a términos más genéricos o específicos gracias a la capacidad de navegación implícita en la estructura jerárquica, a partir de los descriptores y viceversa.

– Los índices permutados (tablas de correspondencias entre descriptores y *CDU*) se generarían y permitirían la búsqueda a partir de los registros introducidos en ambas direcciones: por *CDU* en orden jerárquico (con sus correspondientes descriptores) y por descriptores alfabéticamente por el primer elemento (con su correspondiente notación). En los registros quedaría reflejada esta información en los campos de "Clasificación" y "Descriptores". Así pues, en el primer caso (*CDU*-descriptores), se podría buscar a partir de cada uno de los dos campos todos aquellos registros que tuviesen asociados, ordenándose de más generales a más específicos. En el segundo (descriptores-*CDU*), los grupos de descriptores en línea se ordenarían alfabéticamente por el primero de ellos, quedando a la vez relacionados y próximos teniendo en cuenta que su orden viene determinado por la importancia y novedad de los conceptos de más a menos (orden nuclear).

– Cuando las signaturas incluyan el signo de relación (:) se podrá buscar por los dos elementos. El enlace enviará el usuario a la posición del índice que se correspondería con el segundo elemento facilitando ver otros documentos relacionados: es decir, como opción alternativa de navegación.

– Posibilidad de consultar las notaciones de la *CDU*, destacadas tipográficamente, además de incluir las notaciones.

– Respecto a los descriptores e identificadores, cada uno de los elementos de ambos campos serían vinculados al tesauro y al índice alfabético de palabras clave, respectivamente. En este último caso, las palabras clave utilizadas en la búsqueda que se consideraran como referencias de descriptores aceptados en el tesauro, remitirían a éstos en base a la definición de los registros de autoridad.

– Se podrá navegar por el tesauro alfabéticamente, por orden jerárquico y por facetas que a su vez tendrán en cuenta las subdivisiones de la *CDU* auxiliares generales así como también las específicas.

– Limitar la acción del buscador por palabras clave a determinados campos contribuiría a la recuperación de la información relativamente relevante (en función del sitio dónde el motor o robot buscará). Se trataría de un modelo inicial de búsqueda avanzada que se podría complementar con la búsqueda por campos, opción que excluiría el resumen; el resumen en el primer supuesto (la recuperación por palabra clave) sí que sería objeto de indización por parte de la aplicación.

– Presentar diferentes formatos y versiones de salida de las descripciones: mediante formulario o *ISBD* (para los usuarios), en *MARC* (para el personal bibliotecario) y en versión reducida y detallada (según interés).

– Facilitar el acceso al catálogo vía web mediante una interfaz que se pueda parametrizar y sea fácil de usar.

– Reutilización de las búsquedas.

– Seleccionar de los registros obtenidos como resultado aquellos que interesen para un uso posterior.

– Exportar los resultados.

– Autoprotección frente a las consultas excesivamente largas y evitar la caducidad de las sesiones.

– La no interferencia de los acentos y del uso de mayúsculas/minúsculas.

Conclusiones

Al llegar al final de este trabajo y después de aplicado el sistema de indización que se proponía en una colección representativa de ocho documentos especializados en ciencias sociales, más la experiencia profesional previa, podemos concluir que se ha comprobado su efectividad. En relación a los requisitos sugeridos para el programa informático de gestión, será conveniente que, en colaboración con el personal informático y según el margen o no de presupuesto disponible, se lleven a cabo las conversaciones, demostraciones y pruebas necesarias para establecer las pautas que permitan obtener un sistema lo más eficaz posible respecto a su diseño y a las posibilidades de recuperación de la información.

Queremos destacar, en relación a los procesos de indización y de recuperación, la importancia que adquiere la cadena lógica de descriptores (orden nuclear) al margen de que se puedan lanzar búsquedas a partir de cualquiera de los descriptores e identificadores que la conforman.

Según la bibliografía consultada, faltaría realizar más estudios sobre la elaboración de un formato

Ejemplo de registro y comentarios

Bayés, Ramon and Ernest Garriga. "Repertorios conductuales mínimos en dos grupos de niños de diferente nivel socioeconómico". *Anuario de psicología* 5 (1971): 43-65.

Tipo de documento:
Se trata de un artículo que originariamente se publicó en una revista que editaba dos números al año, hasta 1989 en que la periodicidad se convirtió en trimestral. El documento que analizamos es el resultado de publicar de manera independiente el texto conservando su paginación: se considerará por tanto como separata.

- **Descriptores primarios**

Educación

Enseñanza primaria; Igualdad de oportunidades; Badalona (Catalunya); 1971
371.014.1(467.1 Ba Badalona)"1971"

Educación

Enseñanza primaria; Desigualdad social; Badalona; 1971
371.014.53(467.1 Ba Badalona)"1971"

Educación

Enseñanza primaria; Evaluación inicial; Tests de diagnóstico; Badalona (Catalunya) ;1971
371.263(467.1 Ba Badalona)"1971"

Lenguas

Lengua catalana; Enseñanza primaria; Badalona (Catalunya); 1971
804.99:373.3(467.1 Ba Badalona)"1971"

- **Descriptores secundarios**

Ley General de Educación (1970)
Inmigración; Badalona (Catalunya); 1971

Resumen:

El artículo es el resultado de la investigación realizada en enero y febrero de 1971 por los autores, con el objetivo de hacer un estudio comparativo entre dos muestras de población infantil al iniciarse el ciclo educativo (de 8 niños/niñas de 6-7 años cada una) en relación con el origen familiar, los factores socioeconómicos en el área urbana barcelonesa, concretamente de dos escuelas diferentes de Badalona y en relación al estudio de la lengua catalana. Con motivo de la aprobación de la Ley General de Educación (1970), se pretende demostrar que para garantizar la igualdad de oportunidades para acceder a la Educación General Básica es preciso que todos los alumnos hayan conseguido un repertorio conductual básico de respuesta que en este caso se denomina instrumento BG: aptitudes y habilidades mínimas. Partiendo de la hipótesis que no se cumplen, se explica la metodología utilizada y se hace un análisis estadístico que acaba confirmándola. Se observa también que las diferencias sólo son significativas cuando se trata de las niñas.

Contexto:

– Los autores son profesores e investigadores universitarios catalanes y, en este documento, se complementan en tanto que se aplica la metodología científica en el ámbito de la psicología (aspectos sociales y que condicionan la conducta de los niños durante la etapa de educación primaria habiendo cursado al menos un año de formación preescolar) con los conocimientos estadísticos.

Ramon Bayés es Doctor en Filosofía y Letras (Sección de Psicología) por la Universidad de Barcelona y Diplomado en Psicología Clínica por la misma Universidad. Ha sido profesor. Desde finales de los años 70's se especializó en psicología de la salud y fue desde 1983 catedrático del Departamento de Psicología de la Educación en la Universitat Autònoma de Barcelona.

Ernest Garriga es titular del Departamento de Matemática Aplicada y Telemática de la Universitat Politècnica de Catalunya.

– Desde el punto de vista psicológico es interesante conocer el modelo conductual básico que establecen los autores para evaluar las muestras de población.

– Es importante para el estudio de la historia de la enseñanza primaria en Catalunya y España, por extensión, al final del período franquista. Un año más tarde de la implantación de la Ley General de Educación se describe la población de una ciudad catalana con mucha inmigración procedente del resto del estado y que presentaba diferencias respecto el nivel socioeconómico, de aprendizaje básico, etc. Por lo que tiene que ver con la política educativa del momento también interesa el uso del catalán (esta población no conocía la lengua catalana) y en la educación de las niñas las desigualdades eran más marcadas.

Análisis de la indización

Nos hemos basado en el planteamiento inicial (hipótesis) que se plantean los autores y hemos querido representar la totalidad de conceptos im-

plícitos y explícitos del documento. Hemos definido en la política de indización, que sería una tipología de las que se tratarían a nivel más exhaustivo.

En el caso del concepto 2) se ha optado finalmente por "Desigualdad social" dado que hace referencia al ámbito y situación socioeconómica de los diferentes grupos sociales, mientras que el descriptor de ERIC era más general. Para complementarlo, se ha seleccionado "Igualdad de oportunidades" en tanto que es un término relacionado y el documento trata de las diferencias entre los dos grupos poblacionales desde el punto de vista de su acceso a la educación.

"Tests de diagnóstico", conviene precisar que finalmente nos hemos decidido por la forma en plural del descriptor ya que en el documento se habla explícitamente de un instrumento de análisis concreto elaborado por los autores y denominado "Instrument BG", una descripción del cual se adjunta en un anexo (se trata como materia este instrumento y no el tipo de documento como materia).

Para representar el concepto 3) se prefiere "Evaluación inicial" a "Evaluation utilization" porque es más específico.

"Educación" es el común denominador de la mayoría de descriptores primarios, pues el ámbito de estudio se trata de forma transversal.

Los conceptos 5) y 6) son parciales y sólo hacen referencia a una de las muestras de población estudiadas, por este motivo se han considerado secundarios.

adecuado para el tratamiento de la *CDU*, así como estudiar otros modelos de interfaces que permitan al usuario una consulta más ágil y rápida.

En estos momentos en Europa se está avanzando en el uso de herramientas tradicionales acoplándolas a las nuevas tecnologías, pues ya existen experiencias basadas en el uso de la *CDU* como sistema de recuperación.

Sería deseable y conveniente, así pues, realizar esfuerzos para, esta vez sí, tirar del carro y aprovechar las nuevas tecnologías para innovar y mejorar los sistemas de recuperación existentes.

Notas

1. **Kashyap, Madan-Moham**. "Likeness between Ranganathan's Postulations based approach to knowledge classification and entity relationship data modelling approach". En: *Knowledge organization*, 2003, v. 30, n. 1, pp. 1-19.

2. **Bates, Marcia**. "El context i la interacció han de ser elements bàsics en els sistemes de recuperació de la informació. Entrevista por **Mario Pérez-Montoro** y **Jesús Gascón**". En: *Item*, 2006, n. 42, pp. 51-62.

3. Ver apartado 4.2 donde se hace referencia a la polijerarquía y los registros de la colección de documentos analizados.

Bibliografía

Amato, Carol J. *The world's easiest guide to using the MLA*. Westminster: Stargazer, 1999. ISBN 0-9643853-7-6.

Bates, Marcia. "El context i la interacció han de ser elements bàsics en els sistemes de recuperació de la informació. Entrevista por **Mario Pérez-Montoro** y **Jesús Gascón**". En: *Item*, 2006, n. 42, pp. 51-62.

CDU: Clasificación Decimal Universal. Madrid: AENOR, 1995. ISBN 84-8143-019-6.

Comissió de les Comunitats Europees; Consell d'Europa. Tesaurus europeu de l'educació. Versió en llengua catalana, 2003. Consultado en: 30-06-07.
http://www.doredin.mec.es/documentos/TEECAT.pdf

ERIC (Educational Resources Information Center): processing and reference facility, 2003. Consultado en: 30-06-07.
http://www.eric.ed.gov/

Granados-Colillas, Mariàngels. *Bibliografia sobre Joventut.* Barcelona, 1989.

Granados-Colillas, Mariàngels; **Orrit-Ambrosio, Dionis**. "Informe presentat al Servei de Biblioteques i del Patrimoni Bibliogràfic de la Generalitat de Catalunya corresponent a la revisió del procés d'indexació d'aplicació a l'Hemeroteca Nacional de Catalunya". Barcelona, 16 de desembre de 1991.

Granados-Colillas, Mariàngels; **Nicolau-Payàs, Anna**. "Indexació i resum de 9 documents: aplicació d'un nou sistema d'indexació multilingüe". Barcelona, UB, curs 2003-2004.

Granados-Colillas, Mariàngels; **Nicolau-Payàs, Anna**. "La recuperació de la informació en els catàlegs en línia: l'ús de la Classificació decimal universal i la seva implicació en la indexació". En: *7º Congreso del Capítulo Español de ISKO, 2005*. Barcelona: Universitat de Barcelona. Departament de Biblioteconomia i Documentació, 2005, pp. 249-267.

Houston, James E. *Thesaurus of ERIC descriptors*. 12th ed. Phoenix: Oryx, 1990. ISBN 0-89774-788-7.

Kashyap, Madan-Moham. "Likeness between Ranganathan's Postulations based approach to knowledge classification and entity relationship data modelling approach". En: *Knowledge organization*, 2003, v. 30, n. 1, pp. 1-19.

Lancaster, Wilfrid; **Pinto, María**. *Procesamiento de la información científica*. Madrid: Arco Libros, 2001. ISBN 84-7635-485-1.

Rubió-Balaguer, Jordi, (dir.). *Classificació decimal per a les biblioteques catalanes*. Barcelona: Teide, 1982. ISBN 84-307-7331-2.

San-Segundo, Rosa. "Indización en cadena y su aplicación práctica". En: *La representación y la organización del conocimiento en sus distintas perspectivas: su influencia en la recuperación de la información: actas el IV Congreso ISKO-España*, 1999, pp. 53-57. ISBN 84-8138-435-6.

Slavic, Aida. "UDC implementation: from library shelves to a structured indexing language". En: *Digital library of information science and technology*, v. 33, n. 3, 2004. Consultado en: 30-06-06.
http://dlist.sir.arizona.edu/661/
UDC Consortium. Outline of the UDC. Consultado en: 30-06-06.
http://www.udcc.org/outline/outline.htm

Universal Decimal Classification (BSI Standards). English full ed. London: British Standards Institution, 1975. ISBN 0-580-08626-7.

Mariàngels Granados, Anna Nicolau, *Biblioteca de Catalunya*
mgranados2@gmail.com
anicol@telefonica.net

Normalización de la información: la aportación de *IraLIS*

Por **Tomàs Baiget, Josep-Manuel Rodríguez-Gairín, Fernanda Peset, Imma Subirats** y **Antonia Ferrer-Sapena**

Resumen: *La normalización de la información es imprescindible para transferirla, almacenarla y recuperarla. Afecta a todos los aspectos de su uso, y no sólo a los profesionales de su tratamiento. Este artículo aborda los aspectos más relevantes de la práctica profesional en este campo. Describe la solución implantada para nombres de autores españoles: IraLIS. Concluye insistiendo en la necesidad de que los autores científicos españoles se conciencien de que estas acciones mejoran la visibilidad de sus trabajos.*

Palabras clave: *Normalización, Estandarización, Nombres de autor, Autoridades de nombres personales*

Josep-Manuel Rodríguez-Gairín *(profesor en la Universitat de Barcelona),* ***Imma Subirats*** *(directora del repositorio E-LIS, y técnico de información en la Food and Agriculture Organization, Roma),* ***Tomàs Baiget*** *(responsable de proyectos en el Institut d'Estadística de Catalunya, Barcelona),* ***Fernanda Peset*** *(profesora en la Universidad Politécnica de Valencia) y* ***Antonia Ferrer-Sapena*** *(profesora en la Universidad Politécnica de Valencia). Miembros del Centro Internacional para la Investigación en Estrategia y Prospectiva de la Información (Ciepi).*

Title: Information standardization: the *IraLIS* contribution

Abstract: *Standardization is essential for proper transfer, storage and recovery of information. Standardization affects all aspects of information usage, and not only for professionals in information processing. The most prominent aspects of professional practice in this field are adressed, including the IraLIS solution for citing Spanish authors with multiple surnames. There is a need for Spanish scientific authors to be aware that taking these actions will improve the visibility of their work in international citation resources.*

Keywords: *Standardization, Author names, Personal names citation standards, Personal name authorities*

Baiget, Tomàs; **Rodríguez-Gairín, Josep-Manuel**; **Peset, Fernanda; Subirats, Imma**; **Ferrer, Antonia**. "Normalización de la información: la aportación de IraLIS". En: *El profesional de la información,* 2007, noviembre-diciembre, v. 16, n. 6, pp. 636-643.

DOI: 10.3145/epi.2007.nov.10

Normalización de nombres

LA NORMALIZACIÓN DE LA INFORMACIÓN es imprescindible para transferirla, y afecta a todos los aspectos de la actividad humana, no sólo a los profesionales en el tratamiento de la información.

El trabajo de las entidades de normalización supera todas las barreras nacionales, como por ej., el establecimiento de los nombres de países o sus códigos de la *ISO 3166-1* (1997)[1], las Unidades Territoriales Estadísticas[2] (*NUTS*) utilizadas por la *Comisión Europea*, etc.

En el mundo de la información, según **Herrero-Pascual** (1999), el control de los nombres es una tarea clásica en el entorno bibliotecario, pues afecta directamente a la eficacia de la recuperación de información. Además, ahora se rastrea un mundo infinito de información distribuida por internet y se extienden más las formas de trabajo colaborativas y virtuales (**Ontalba**, 2002 y 2006).

Tradicionalmente las bibliotecas han aceptado en sus ficheros de autoridad una forma del nombre personal o corporativo de un autor, haciendo referencia a él desde las formas no aceptadas. La *IFLA* promovió la edición de unas directrices para las entradas de autoridad y referencia, conocidas como *GARE* (*Guidelines for authorities and reference entries*), publicadas en 1984 (1993 en España) que abarcan los encabezamientos de autor personal, autor corporativo y títulos uniformes. Por otra parte, las directrices para los registros de autoridad y referencia de materias (*Guidelines for subject authority and reference entries*, *GSARE*), editadas en 1993 y traducidas en 1995, abarcan las entradas de materia y el lugar geográfico.

La segunda edición (2001) de las *GARE* (siglas desglosadas esta vez como *Guidelines for authority records and references*) —en las que intervino **Assumpció Estivill** como representante española-, fueron traducidas por la *Dirección General del Libro, Archivos y Bibliotecas* del *Ministerio de Cultura*, teniendo en cuenta las más recientes recomendaciones en cuanto a nivel mínimo para registros de autoridad (*Minimum level authority record*, *MLAR*) que recoge un número, su lengua, la autoridad, entidad,

referencias y notas), número normalizado (*International standard authority data number*, *Isadn*) y requisitos funcionales de los registros bibliográficos (*FRBR*).

Tillet (1996) nos recuerda que el *Isadn* es propuesto por la *IFLA* desde 1980, siendo uno de los éxitos de su programa *Universal bibliographic control and international MARC* (*Ubcim*). Hace unos años también los archiveros han editado las *International standard archival authority record (for corporate bodies, persons and families) o Isaar (CPF)*.

Los objetivos básicos de las *GARE* y las *GSARE* fueron definir los elementos necesarios en los registros de autoridad, asignarles un orden y establecer su estructura. Seguían las directrices del programa *Control bibliográfico universal* (*CBU*) en que las agencias nacionales se responsabilizaban de elaborar las listas de autoridades.

Pero actualmente la situación ha cambiado y se hace necesario un fichero virtual de autoridades a escala internacional que contemple diferentes formas de escritura pero un único acceso. Además los nuevos modelos conceptuales como *Functional requirements for bibliographic records* (*FRBR*) o *Functional requirements and numbering of authority records* (*Franar*) indican que el usuario lo que necesita es identificar sin ambigüedad, encontrar los conceptos, y acceder al item.

El control de la descripción ahorra tiempo y esfuerzo al usuario, mostrándole referencias de las entradas autorizadas. Ordena los registros bajo una sola forma, con indicación de las no aceptadas y permite generar automáticamente las variantes (orden directo, permutado, abreviado, otras lenguas y escrituras…). Para **Tillet** estos cambios vienen determinados por la aparición del web o más recientemente la *Open archives initiative* (*OAI*), que hacen posible una navegación por un universo casi infinito de recursos y precisan de un control persistente a la vez que interoperable entre los diversos formatos. Ya hoy la Oficina de desarrollo en red y estándares MARC de la Library of Congress ha elaborado un esquema de metadatos para autoridades en XML (*Metadata authority description schema*, *MADS*). Incluye los nombres de personas, de organizaciones, de congresos y eventos, de términos, de nombres geográficos…, para compartir en internet este tipo de información.

Estos nuevos escenarios, necesarios en el trabajo con autoridades, se concretan en diversos proyectos, algunos financiados por la Unión Europea, como *Author*, descrito por **Danskin** (1996). En él se realiza una equivalencia de los registros de autoridad de siete países al mismo formato de comunicación: *Unimarc*. También la *OCLC* a través de *Connexion* (antes *CORC*) busca la interoperabilidad, presentando los registros en MARC21 y *Dublin core*. Otros, como *Linking and exploring authority files* (*LEAF*), intentan enlazar los ficheros de autoridad usando los protocolos Z39.50 y *OAI*. Por su parte, los proyectos relacionados *<indecs>* e *Interparty* promueven la cooperación entre bibliotecas, archivos y museos. Por último la *Hong Kong Chinese authority for names* (*Hkcan*), proporciona un fichero de autoridades compartido entre las bibliotecas de su consorcio, permitiendo la forma oriental, transcrita y simplificada.

El origen de estos esfuerzos puede verse en la sección *Name authority cooperative program* (*NACO*) del *Program for cooperative cataloging* (*PCC*) que nació en 1976 con un cariz casi exclusivamente anglosajón. Hoy día, no obstante, permite la participación de otras bibliotecas si son miembros del citado *PCC*. **Byrum** (2003) cifraba su oferta en dos millones de registros, con un aumento anual de 200.000.

Para **Tillet**, aplicar un control de autoridades actualmente permite mayor precisión en la recuperación, algo casi imprescindible en un entorno web. Además, la estructura de los registros de autoridad permite navegar entre múltiples recursos de forma transparente para el usuario entre información diferente referida a una autoridad: nombre, explicación de variaciones, incoherencias, biografías, fuentes… (**Weber**, 2002, p. 5):

"It is widely accepted that the national and international sharing of authority information is a suitable means for reducing the costs of cataloguing work in libraries and archives as well as the costs for biographical research work undertaken in scientific projects".

El nombre de un autor personal, a diferencia de los nombres de las instituciones –o autores corporativos– son independientes del idioma, especialmente si no necesitan transliteración. Al fin y al cabo son cadenas de caracteres unidas sin significado, y por tanto no son susceptibles de traducción o variación. Por ello resultan tan fáciles de recuperar de manera eficiente, al igual que la mayoría de nombres de empresas. Por ej., cuando alguien introduce en *Google* un nombre de persona o de una empresa suele obtener la información: de la empresa muy eficientemente porque usualmente tendrá un sitio web propio que aparece el primero en el ranking de resultados. Pero los autores, al figurar con variantes de nombres –nombre y apellido, apellidos y nombre, nombre desarrollado, con iniciales…- la labor de discriminación es ardua según el número de páginas que devuelva la consulta. El problema de los nombres es de singular importancia para el personal académico, que basa su reconocimiento profesional en su currículo personal e institucional.

Problema añadido para los autores españoles

Las variaciones en las firmas cobran especial relevancia en España y en los países hispanoamericanos. A diferencia de los países anglosajones, portugueses, nórdicos y eslavos, usamos primero el apellido del padre y luego el de la madre, y muchas veces también un nombre de pila compuesto. El hecho de que una firma esté compuesta por muchos elementos contribuye, además, a que el autor firme sus trabajos de distintas formas. En España, por añadidura, también contamos con las variantes derivadas de las traducciones del nombre a los diferentes idiomas locales, adición de partículas, los apelativos o diminutivos… Los autores que firman con todo su nombre oficial (tal como figura en su DNI, pasaporte, etc.) se encuentran a menudo con la desagradable sorpresa de que sus trabajos publicados en revistas científicas aparecen referenciados de diferentes formas en los buscadores, en los depósitos OAI y en las bases de datos internacionales. Según como el productor del recurso haya interpretado su nombre aparecerán nuevas variantes, por lo que recopilar su bibliografía personal es en ocasiones muy difícil. En este contexto, la *Fundación Española para la Ciencia y la Tecnología* (*Fecyt*) publicó en enero de 2007 unas Recomendaciones (2007) para la firma de autores personales. El estudio de **Ruiz**, **Delgado** y **Jiménez** (2002) sobre las bases de datos *ISI*, *Medline* e *IME*, permite identificar las formas en que los autores españoles firman y cómo aparecen reseñados.

Así, estas recomendaciones y el proyecto que ahora presentamos, *IraLIS*, mantienen una estrecha relación en sus principios. Existen, no obstante, dos diferencias. *IraLIS* es un banco de prueba para los autores de un área del conocimiento, biblioteconomía y documentación, muy sensible a este tipo de cuestiones de normalización, circulación, visibilidad, impacto… Por otra parte, tiene un alto componente tecnológico, desarrollado por **Josep-Manuel Rodríguez-Gairín** (*Universitat de Barcelona*), lo que permite obtener unos resultados directamente útiles para los autores científicos en las cuestiones apuntadas.

La aportación de IraLIS

IraLIS (*International Registry for Authors in Library and Information Science*) es un proyecto que surge en noviembre de 2006 a partir de las relaciones entre *E-LIS* (*Eprints in library and information science*), *EXIT* (*Directorio de expertos en el tratamiento de la información*) y la revista *EPI* (*El profesional de la información*).

Respectivamente *E-LIS* es un repositorio especializado en ciencias de la documentación, que cumple las normas *OAI-PMH*; *EXIT* es un directorio internacional con información sobre personas, instituciones y especialidades de trabajo; y por último, *EPI* es una revista española de nuestra profesión, la única que es indizada en *ISI-Thomson* y *Scopus*. Estos tres productos tienen en común su proyección internacional, su carácter documental, su distribución por internet, y el estar orientados al mundo de la gestión de la información.

http://www.iralis.org
http://eprints.rclis.org
http://www.directorioexit.info/
http://www.elprofesionaldelainformacion.com

Desde *EXIT* se recuperan los trabajos depositados en *E-LIS* a través de la interrogación de los nombres de los autores. De este vínculo surge la necesidad de trabajar en la normalización de los nombres de autoridades, algo que ya *EPI* había detectado al analizar la circulación de los trabajos de sus autores en las bases de datos internacionales.

El hecho de que todos estos productos sean complementarios entre sí y estén diseñados por personas vinculadas al mundo de la información determinó el comienzo de *IraLIS* para registrar, recomendar y recuperar las formas diversas en que un autor puede firmar. Su estrategia de normalización resulta ecléctica si la comparamos con el trabajo tradicional de catalogación de autoridades que hemos expuesto. Sin embargo, no deja de estar inmerso en este contexto, trabajando con autores personales y también corporativos –en tanto que calificadores de los autores personales-, como veremos más adelante. En definitiva es un desarrollo tecnológico que intenta superar los silencios que se producen en la recuperación de información sobre autores en internet. De sus resultados se beneficiarán en última instancia los proyectos mencionados y los autores que obtengan un iralis, es decir, el nombre registrado para su firma.

Objetivos de IraLIS

IraLIS se ha fijado unos objetivos a corto plazo, y otros a medio-largo plazo, pues las tareas a realizar requieren tanto acciones contundentes inmediatas (cuando están en la mano de sus promotores), como actuaciones más pausadas pero continuas de penetración en el entramado científico.

Se propone reducir la grave distorsión en la recopilación bibliográfica de los autores de tres formas:

1. Creando un registro de nombres de autores en biblioteconomía, documentación y archivística, que ayude a resolver las diferentes variantes. El registro incorporará tanto las que puede haber usado un autor, como las que haya interpretado el productor, agregador, buscador, etc., de las diversas fuentes de información.

2. Concienciando a los autores hispanos para que firmen sus trabajos siempre de la misma forma, pensando en cómo los referencia-

rán las bases de datos internacionales, los archivos *OAI* y los robots de búsqueda. Sistemas que, como hecho consumado prácticamente irreversible, están bajo la influencia de la cultura y los hábitos ingleses (*Science Citation Index*, *Scopus*, *Chemical Abstracts*, *Medline*, *Google Scholar*, etc.).

3. Creando el sencillo formato de firma *IraLIS*, que permite ser interpretado adecuadamente y sin confusiones también por las fuentes de información de cultura anglosajona. A partir de esta forma aceptada (o iralizada) del nombre se pretende que un algoritmo de consulta en motores de búsqueda y *harvesters* permita recuperar todas las variantes.

Así, *IraLIS* no es únicamente un registro de la forma estandarizada del nombre, sino que se basa en la interoperabilidad de los sistemas y en la recuperación del nombre del autor desde diferentes bases de datos abiertas. Por ejemplo, *IraLIS* sabe contestar en XML a preguntas hechas en *OpenURL*, y el campo iralis del directorio *EXIT* muestra de forma dinámica los datos que están registrados en *IraLIS*. Esta funcionalidad permitirá igualmente que desde repositorios como *E-LIS* pueda validarse la introducción de autores mediante consultas directas a *IraLIS* usando tecnologías *ajax*.

Funcionamiento

El proyecto cuenta con un comité ejecutivo, formado por los miembros fundadores, que revisa la pertinencia de los iralis registrados para hacerlos visibles o no. Además, cuenta con un comité asesor que propone nuevas experiencias en este campo, sugiere mejoras o son consultados en caso de dudas. Actualmente forman parte de él representantes de organismos relacionados con la gestión y el estudio de la ciencia española, como son la *Fecyt*, varias universidades, el *Cindoc* o el *Parc de Recerca Biomèdica de Barcelona*[3].

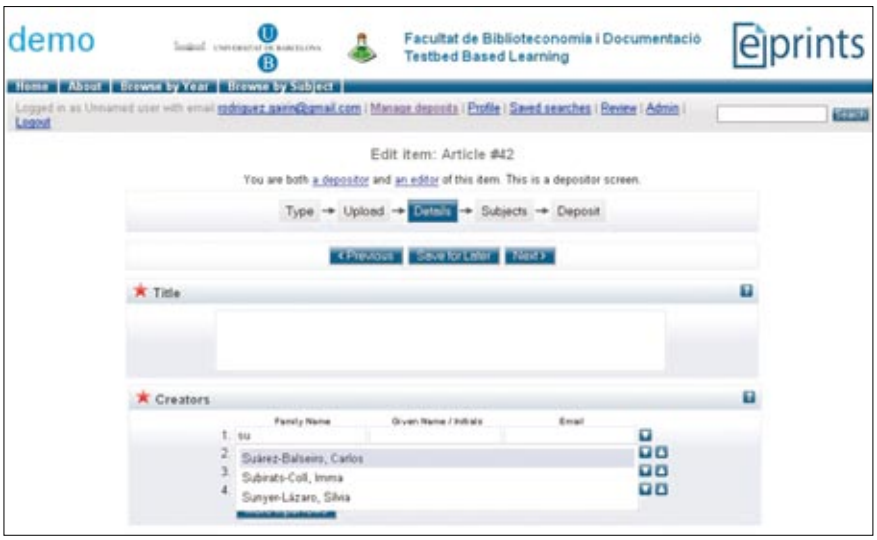


Figura 1. E-prints. Demo. http://eprints3.bd.ub.es/

http://www.iralis.org/?q=node/11

La programación de *IraLIS* ha dado lugar a una base de datos o lista de autoridades, cuyos registros son generados por los propios autores rellenando un sencillo formulario.

Figura 2. Register your IraLIS. http://www.iralis.org/?q=node/8

Una vez introducido el nombre completo, el sistema presenta una lista con las diferentes variantes que pueden constituir el iralis personal en forma inversa. Así, si un usuario introduce por ejemplo el nombre Roberto **José Andrade Pallarés,** el sistema le ofrece esta gama de opciones para que de entre ellas el usuario elija el iralis que prefiera.

A sugerencia de **María Bordons** (*Cindoc*) se omitieron algunas variantes de manera que se trate de una selección de posibilidades, no una lista exhaustiva:

Uno de estos formatos iralis-compatibles, que como se puede apreciar siempre están constituidos por 2 ristras de caracteres (con una eventual inicial intermedia opcional), debería ser el formato de firma fijo y universal de cualquier autor de trabajos científico-técnicos, susceptibles de ser recogidos por las bases de datos y los robots de búsqueda. En términos generales coincide con las Recomendaciones de la *Fecyt*, ya nombradas, usando guiones o combinando Nombre1 Nombre2 Apellido1Apellido2.

Por último, *IraLIS* ofrece una recomendación de firma e información sucinta sobre el funcionamiento de las bases de datos bibliográficas. Una vez registrado, el sistema podrá utilizar tanto las variantes definidas por el autor como las au-

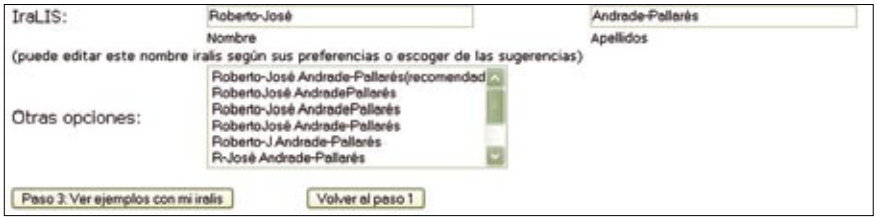


Figura 3. El sistema ofrece variantes normalizadas para que el interesado elija la que prefiere. La forma sugerida que se presenta en la primera línea puede editarse.

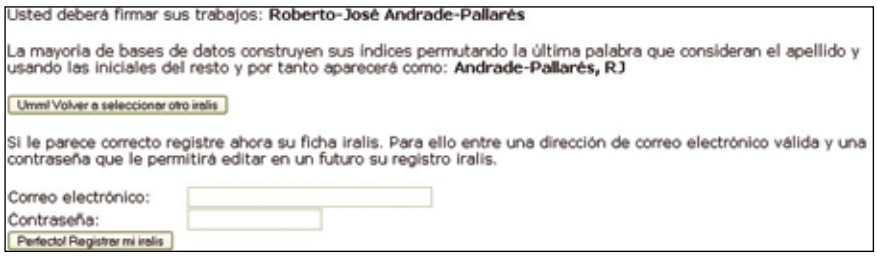


Figura 4. Registro definitivo del formato de firma elegido

togeneradas por *IraLIS*, junto con el correo electrónico del autor –a sugerencia de **Isidro Aguillo**, *Cindoc*- para la recuperación de información por internet.

Una vez registrado el iralis se puede ver en formato *MADS* y definir las variantes de nombre. Así también, presenta un enlace al *Directorio de Expertos en el Tratamiento de la Información* (*EXIT*) que permite completar el registro con datos de tipo directorio.

Sin embargo puede ocurrir que en publicaciones oficiales no sea posible firmar con el iralis, o también puede ser que el autor no quiera usar ningún formato iralis. En estos casos el registro *IraLIS*, con todas las variantes, tanto las reales como las potenciales, seguirá siendo de gran utilidad cuando se quieran hacer búsquedas exhaustivas de las obras de un autor.

Estas reglas de validación del proyecto son claras: el nombre ha de estar compuesto por dos ristras que forman nombre y apellido. No aparecerán visibles en el proyecto las personas que deseen mantener otro tipo de firma. Aunque se permite el registro de cualquier formato de firma, el Comité Ejecutivo de *IraLIS* puede autorizar o no las formas introducidas según la regla de validación expuesta: dos únicas ristras de caracteres.

Formato de los registros

El registro de un autor está compuesto por:

ID_iralis: Un número
Fecha de alta: fecha de entrada en el registro
Nombre registrado: con un campo para los apellidos reales y un campo para el nombre real. Recoge la forma del nombre
Iralis registrado: con un campo para los iralis-apellidos y un campo para el iralis-nombre. Recoge la forma aceptada.
Correo electrónico: usado por el autor
Password

Las variantes del nombre contienen los siguientes campos:

Id_iralis
Variante_nombre
Variante_apellidos
Cronología: con los campos Desde y Hasta,
Comentarios

Cada autor posee un número iralis. Este identificador se autogenera con los datos que se poseen en la siguiente forma XXZZZ00000, donde XX es el código del país, ZZZ es la especialidad (en previsión de la extensión del proyecto a otras disciplinas) y 00000 el número automático del registro de *IraLIS*. Así por ejemplo el ID de Tomàs Baiget es ESLIS00010.

Para normalizar los números de identificación los ficheros de autoridades que generan, por ejemplo, las bibliotecas nacionales se utiliza el *International standard authority data number* (*Isadn*). Pero existen otros números relacionados con grandes bases de datos como puede ser *Scopus*, de *Elsevier*, como la figura que mostramos:

O los de la *Biblioteca del Congreso* de Washington (figura 7).

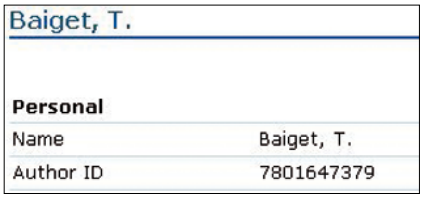


Figura 6. http://www.scopus.com/scopus/author/profile.url?aid=7801647379&origin=recordpage&txGid=W4bRv0OfFuhxm6jrbBXFS_U%3a20

La forma aceptada con nombre y apellido. Es la forma del nombre de un autor, aceptada por él mismo. El autor podrá modificarla en cualquier momento que lo desee, almacenándose la forma antigua para poder seguir recuperándola. Siempre que cumpla la regla de validación aparecerá como aceptada y visible en el sistema. El formato iralis se concibe en forma directa, no anteponiendo apellido, aunque el sistema puede mostrarlo invertido según convenga, como por ejemplo para ofrecer las variantes, para

Figura 5. Ficha de autor con todas las variantes, históricas y actuales, de las firmas que ha usado o han interpretado quienes le han citado

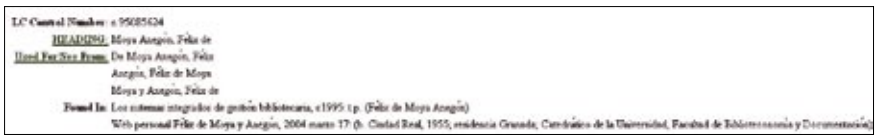


Figura 7. Library of Congress Authority. http://authorities.loc.gov/cgi-bin/Pwebrecon.cgi?AuthRecID=1293991&v2=1&HC=1&SEQ=20070606112309&PID=28270

recuperarlas, para ordenarlas, para visualizarlas, etc.

Variantes con sus explicaciones. Además de las variantes propuestas automáticamente por el sistema, cada autor podrá introducir sus propias variantes, siempre que cumplan las reglas de validación de *IraLIS*. Así, por ejemplo, un autor puede tener un nombre determinado en el Registro Civil de su país, pero él puede decidir firmar con una variante iralis de su nombre: José, Xose, Josep, Joseph, Pepe, etc.

Las variaciones del nombre las puede introducir el mismo autor, indicando en qué momentos utilizó ese nombre y las razones. Pero también podrían ser volcadas desde los ficheros de autoridad, por ejemplo de la *Biblioteca Nacional de España* si su catálogo de autoridades se pudiera interrogar en XML. Consideramos que el porcentaje de autores que utilicen *IraLIS* y tengan entrada en la *Biblioteca Nacional* va a ser reducido, pues aparecen sólo los que cuentan con una publicación monográfica. Se mostraría en la ficha como una recopilación de las variantes.

Además el formato iralis es compatible con el esquema de descripción *Metadata authority description schema* (*MADS*), que transforma el MARC en xml, como muestra la siguiente figura 9.

Y todo esto para conseguir recuperar la bibliografía de los autores exhaustivamente. Actualmente ya se encuentran implementados algoritmos de búsqueda de la forma aceptada y sus variantes en *Google*, *Google Scholar* y *OAIster*. Puede apreciarse en la imagen la diferencia de recuperación para un mismo autor:

En el caso de un motor de búsqueda recupera páginas web que donde aparece el nombre del autor. En el caso de bases de datos bibliográficas abiertas o *harvesters* como *OAIster* devuelve como respuesta referencias bibliográficas:

Google Scholar indiza los nombres de los autores en muchos casos por el último apellido (en nuestro ejemplo, Falgueras EA), por lo que no sería recuperable por el primero. Esto *Google* lo solventa porque indiza el texto completo normalmente el nombre correcto del autor es citado en el texto.

```
- <mads xsi:schemaLocation="http://www.loc.gov/mads/ http://www.loc.gov/standards/mads/mads.xsd">
  - <authority ID="ESLIS0071">
    - <name>
        <namePart type="family">Ferrer</namePart>
        <namePart type="given">Antonia</namePart>
      </name>
    </authority>
  </mads>
```

Figura 9. http://www.iralis.org/? q=node%2F8&paso=10&letra=F&id=71

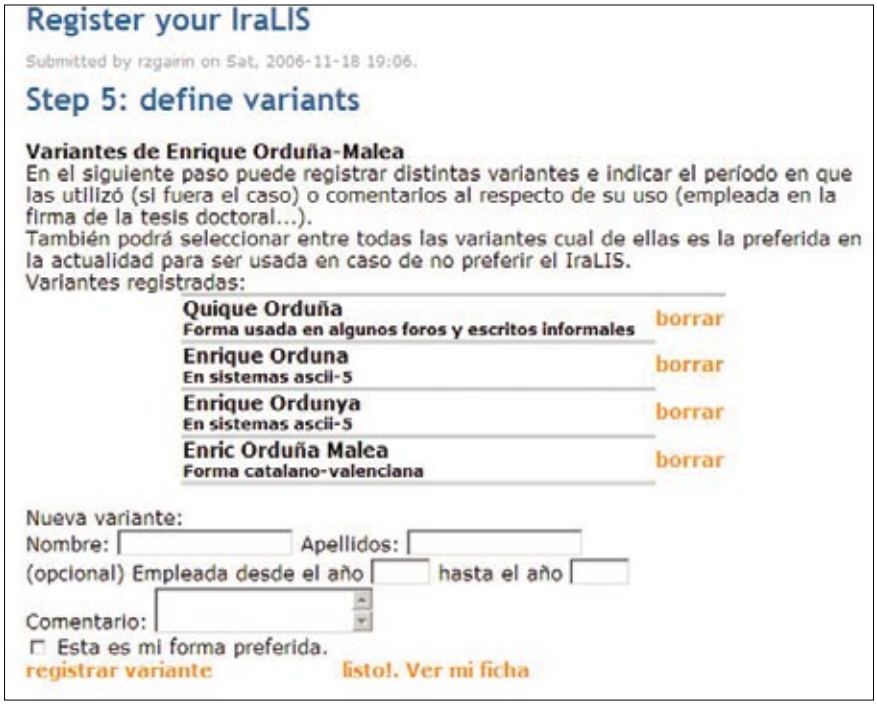


Figura 8. Form para introducir y definir variantes personales

Conclusiones

Hay que resaltar dos aspectos del proyecto *IraLIS* en cuanto a la normalización de la información y su posterior recuperación. En primer lugar, *IraLIS* es un sistema que ayuda a mentalizar a la gente para que firme siempre igual, que recomienda unas formas de firmar y que, en el futuro, recogerá todas las posibles variantes de las firmas de un autor, etc.

En segundo, "iralis" es un nombre de autor en un formato determinado, escogido por un autor de entre diferentes variantes, todas las cuales cumplen ser formatos iralis (dos bloques de caracteres). Por ejemplo tanto MF Peset como Fernanda Peset como Maria-Fernanda

Figura 10. Búsqueda de un autor en Google Scholar (http://scholar.google.es/) a partir de su registro en IraLIS

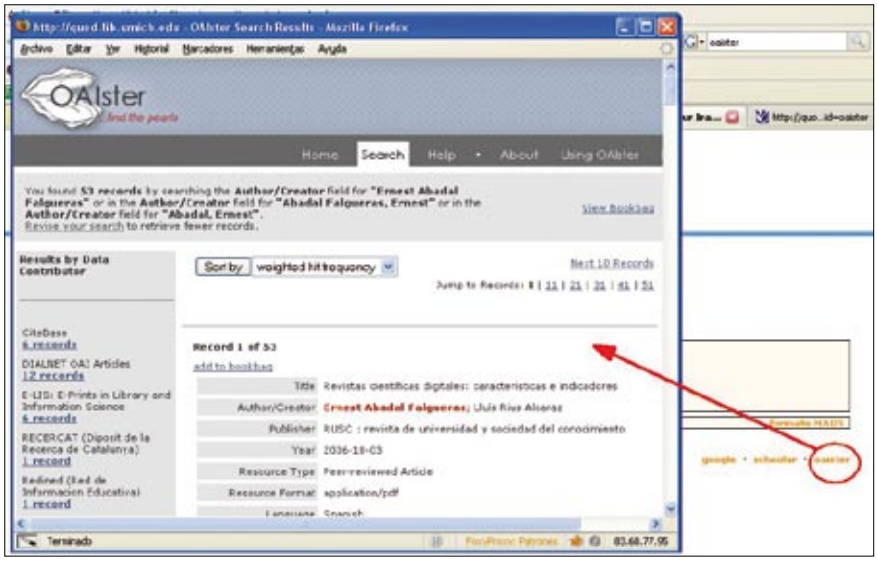


Figura 11. Búsqueda de un autor en OAIster (University of Michigan, Digital Library Production Service). http://www.oaister.org

Peset-Mancebo podrían serlo. El autor escogerá una de esas variantes y la pondrá como su iralis.

Este sistema tiene dos utilidades principales: permitirá la recuperación exhaustiva de bibliografías personales gracias a los registros del sistema *IraLIS*. A través de sus registros pueden recuperarse las variantes en cualquier base de datos. Estas variantes estarían constituidas por todas las formas no escogidas como iralis por el autor al registrarse en el sistema, el correo electrónico y todas las formas que desee introducir un autor, ya sea por que durante un tiempo firmara así o por que se ha encontrado sus trabajos registrados en esa forma en las bases de datos. Desde *IraLIS* se podrán lanzar búsquedas contra bases de datos y buscadores de internet.

Gracias a la normalización y la conexión con las otras bases de datos, *E-LIS* y *EXIT*, facilitará la elaboración de estudios estadísticos de tipo bibliométrico, indicadores socio-profesionales, por género, etc.

En definitiva, *IraLIS* es un prototipo orientado a las ciencias de la documentación y la información que permite ser banco de pruebas para extender su programación y la metodología de difusión a todos los campos de las ciencias en España.

Notas

1 Además a partir de la reunión en Santiago de Compostela en mayo de 2007 *ISO/TC 46/WG 2* se consideró la posibilidad de trabajar en el código de nombre océanos y mares, países históricos, organizaciones internacionales…

2. Nomenclatura de las Unidades Territoriales Estadísticas utilizadas por la Unión Europea

3. Los actuales integrantes de Comité Asesor de *IraLIS* son: **Isidro F. Aguillo** (*Cindoc*, *CSIC*), **Txema Báez** (*Fecyt*), **María Bordons** (*Cindoc*, *CSIC*), **Jordi Camí-Morell** (*Parc de Recerca Biomèdica de Barcelona*), **Emilio Delgado-López-Cózar** (*Universidad de Granada*), **Assumpció Estivill** (*Universitat de Barcelona*), **Isabel Gómez-Caridad** (*Cindoc*, *CSIC*), **Vicente P. Guerrero-Bote** (*Universidad de Extremadura*), **Félix de Moya-Anegón** (*Universidad de Granada*) y **Elías Sanz-Casado** (*Universidad Carlos III de Madrid*)

Bibliografía

Bennett, Rick; **Hengel-Dittrich, Christina**; **O'Neill, Edward T.**; **Tillett, Barbara B.** "VIAF (Virtual international authority file): Linking die Deutsche Bibliothek and Library of Congress name authority files". En: *World library and information congress: 72nd IFLA general conference and council 20-24 August 2006, Seoul, Korea*. Consultado en: 26-10-2007. *http://www.ifla.org/IV/ifla72/papers/123-Bennett_trans-es.pdf http://www.ifla.org/IV/ifla72/papers/123-Bennett-en.pdf*

Buizza, Pino; **Gerrini, Mauro**. "Control de los puntos de acceso de autor y título". En: *Principios de catalogación de IFLA: Pasos hacia un Código Internacional de Catalogación: Informe de la 1ª reunión IFLA de expertos sobre un Código internacional de catalogación, Frankfurt, 2003*. Editado por **Barbara B. Tillet**, **Renate Gömpel** y **Susanne Oehlschläger**. Madrid: Subdirección General de Coordinación Bibliotecaria, 2005, pp. 121-151. Consultado en: 26-10-2007. *http://travesia.mcu.es/documentos/principios_catalogacion_ifla.pdf*

Byrum, John D. "NACO: a cooperative model for building and maintaining a shared name authority database". En: *International conference authority control: definition and international experiences,* Florence, Italy, February 10-12, 2003. Consultado en: 26-10-2007. *http://www.sba.unifi.it/ac/relazioni/byrum_eng.pdf*

Callon, Michel; **Courtial, Jean Pierre**; **Penan, Hervé**. *Cienciometría: la medición de la actividad científica, de la bibliometría a la vigilancia tecnológica*. Gijón: Trea, 1995.

Danskin, Alan. "International standards in authority data control: costs and benefits". En**:** *62nd IFLA General Conference - Conference Proceedings - August 25-31, 1996.* Paris: Inist, 1996. Consultado en: 22-06-2007. *http://www.ifla.org.sg/IV/ifla62/62-dana.htm*

Figura 12. Entre las fichas de IraLIS y del Directorio EXIT existen enlaces en ambas direcciones

Estivill, Assumpció; **Abadal, Ernest**; **Frangannillo, Jorge**; **Gascón, Jesús**; **Rodríguez-Gairín, Josep-Manuel**. "Uso de metadatos Dublin Core en la descripción y recuperación de artículos de revista digitales = Use of Dublin Core metadata for describing and retrieving digital journals". En: *Proceedings DC-2005 International conference on Dublin Core and metadata applications, 12-15 September. Madrid.* Consultado en: 26-10-2007.
http://eprints.rclis.org/archive/00004792/

Estrategia nacional de ciencia y tecnología: Comisión Interministerial de Ciencia y Tecnología. Madrid: Fecyt, 2007.

Fecyt. *Recomendaciones para la correcta identificación de las publicaciones científicas*. Propuesta de manual de ayuda a los investigadores españoles para la normalización del nombre de autores e instituciones en las publicaciones científicas, 17 enero 2007.
http://www.accesowok.fecyt.es/recomendaciones_publicaciones.html

Formato Ibermarc para registros de autoridades. Madrid: Biblioteca Nacional, 1999.

Heijligers, Ton. "¿'Entrada principal' en el futuro?" En: *Principios de catalogación de IFLA: Pasos hacia un código internacional de catalogación: Informe de la 1ª reunión IFLA de expertos sobre un código internacional de catalogación, Frankfurt, 2003*. Editado por **Barbara B. Tillet**, **Renate Gömpel** y **Susanne Oehlschläger**. Madrid: Subdirección general de coordinación bibliotecaria, 2005, pp. 153-158. Consultado en: 26-10-2007. *http://travesia.mcu.es/documentos/principios_catalogacion_ifla.pdf*

Herrero-Pascual, Cristina. "El control de autoridades"**.** En: *Anales de documentación*, 1999, n. 2, pp. 121-136. Consultado en: 14-3-2007.
http://www.um.es/fccd/anales/ad02/AD08-1999.PDF

IFLA. Directrices para los registros de autoridad y referencia de materia. Madrid: Anabad - Arco Libros, 1995.

IFLA. Directrices para los registros de autoridad y referencia. Madrid: Anabad - Arco Libros, 1993.

IFLA. Directrices para registros de autoridad y referencias. Ministerio de Cultura, 2004. Consultado en: 26-10-2007. *http://travesia.mcu.es/documentos/directrices_autoridad.pdf*

Indicadores bibliométricos de la actividad científica española (1990-2004). Madrid: Ministerio de Educación y ciencia, Fundación Española para la Ciencia y la Tecnología, 2007.

Informe SISE 2006: Sistema integral de seguimiento y evaluación, Observatorio permanente del sistema español de ciencia-tecnología-sociedad. Madrid: Fecyt, 2007.

ISO 3166-1 Codes for the representation of names of countries and their subdivisions. Part 1: Country codes. Ginebra: ISO, 1997.

LEAF (Linking and exploring authority files). *Public progress LEAF's second year. Report 2*.
http://www.leaf-eu.org

Mandatory data elements for internationally shared resource authority records: Reports of the IFLA Ubcim working group on minimal level authority records and Isadn. Frankfurt: IFLA Ubcim Programme, 1998. Consultado en: 26-10-2007.
http://www.ifla.org./VI/3/p1996-2/mlar.htm

Metadata authority description schema (MADS). Consultado en: 26-10-2007.
http://www.loc.gov/standards/mads/mads-outline.html

Ontalba-Ruipérez, José-Antonio. "Las comunidades virtuales académicas y científicas españolas: el caso de RedIris". En: *El profesional de la información*, 2002, sept.-oct., v. 11, n. 5, pp. 328-338.

Ontalba-Ruipérez, José-Antonio. "Las comunidades virtuales como herramientas para la socialización del conocimiento tácito". En: *Documentación digital*. Universidad Pompeu Fabra, 2006.

Peset, Fernanda; **Ferrer, Antonia**; **Baiget, Tomàs**; **Rodríguez-Gairín, Josep-Manuel**. "DSpace and the standardization of the information: names of Spanish authors". En: *DSpace user group meeting 2007, FAO, Rome, Italy, October 17th-19th.*

Recomendaciones para la correcta identificación de las publicaciones científicas. Madrid: Fecyt, 2007. Consultado en: 10-03-2007. *http://www.accesowok.fecyt.es/recomendaciones_publicaciones.html*

Reglas de catalogación. Edición refundida y revisada. Madrid, Dirección General del Libro, Archivos y Bibliotecas, 1995.

Ruiz-Pérez, Rafael; **Delgado López-Cózar, Emilio**; **Jiménez-Contreras, Evaristo**. "Spanish personal names variations in national and international biomedical databases: implications for information retrieval and bibliometric studies". En: *Journal of the medical library association,* 2002, n. 90, pp. 411-430. Consultado en: 15-05-2007.
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=128958

Tillett, Barbara B. "Authority control: State of the art and new perspectives". En: *International conference authority control: definition and international experiences,* Florence, Italy, February 10-12, 2003. Consultado en: 26-10-2007. *http://www.sba.unifi.it/ac/relazioni/tillett_eng.pdf*

Tillett, Barbara B. "Un fichero de autoridades internacional virtual". En: *Principios de catalogación de IFLA: Pasos hacia un Código internacional de catalogación: Informe de la 1ª reunión IFLA de expertos sobre un código internacional de catalogación, Frankfurt, 2003.* Editado por Barbara B. Tillet, Renate Gömpel y Susanne Oehlschläger. Madrid: Subdirección general de coordinación bibliotecaria, 2005, pp.95-107. Consultado en: 26-10-2007. *http://travesia.mcu.es/documentos/principios_catalogacion_ifla.pdf*

Weber, Jutta. "Malvine, LEAF and Kalliope: Some co-operation models". En: *Digital access to book trade archives (Papers of the 2001 Conference in The Hague)*. Leiden: Academic Press, 2002. pp. 49-68. Consultado en: 4-3-2007.
http://www.malvine.org/malvine/publications/MALVINE_Weber_MLK.pdf

***Tomàs Baiget**, Institut d'Estadística de Catalunya (Idescat)*
baiget@sarenet.es

***Josep-Manuel Rodríguez-Gairín**, Universitat de Barcelona*
rodriguez.gairin@ub.edu

***Fernanda Peset**, Universidad Politécnica de Valencia*
mpesetm@upv.es

***Imma Subirats**, Food and Agriculture Organization (FAO), Roma*
imma.subirats@gmail.com

***Antonia Ferrer-Sapena**, Universidad Politécnica de Valencia*
anfersa@upv.es

FUNDACIÓN
Alonso
Quijano

para el fomento de la lectura

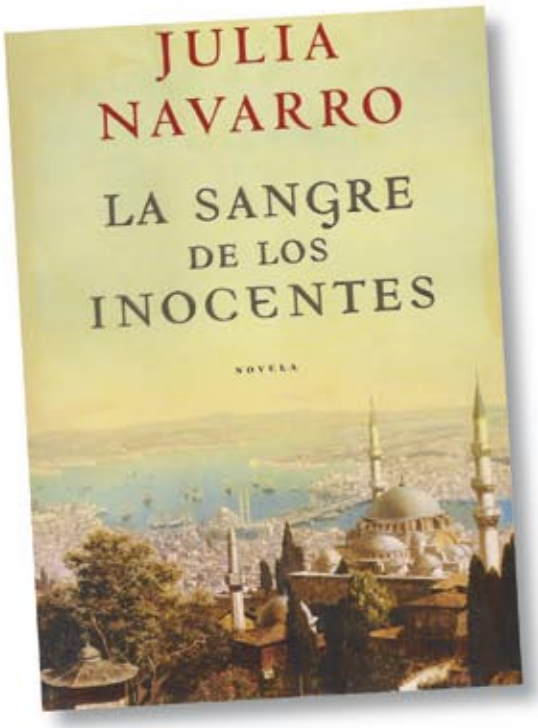
¿qué hacemos?

Fomento de la lectura con menores hospitalizados.
Actividades de Formación.
Revista *Mi Biblioteca.*
Cooperación Internacional.
Recursos sobre lectura en la web.

¿quieres colaborar?

Hazte socio/a y recibirás estos dos libros de regalo*.

Cuota mínima: 20€ al año.



* Promoción válida hasta agotar existencias.

Puedes hacerlo por teléfono: **902 362 869 - 952 23 54 05**
o a través de nuestra web: **www.alonsoquijano.org**

Asóciate y disfruta de estas ventajas:

- Regalo de un libro cada año como agradecimiento por la colaboración con la Fundación.
- Información sobre las actividades de la Fundación y participación en sorteos y promociones.
- Descuentos en suscripciones a revistas del sector.
- Descuentos en los cursos y otras actividades de formación organizadas por la Fundación.
- Regalo del *Calendario de la Lectura* que la Fundación publica cada año.
- Ventajas fiscales según la legislación vigente sobre mecenazgo.

SCImago journal & country rank: un nuevo portal, dos nuevos rankings

Por **Grupo Scimago**

Grupo Scimago. "*SCImago journal & country rank:* un nuevo portal, dos nuevos rankings". En: *El profesional de la información,* 2007, noviembre-diciembre, v. 16, n. 6, pp. 645-646.

DOI: 10.3145/epi.2007.nov.11

EL PORTAL *SCIMAGO JOURNAL & COUNTRY RANK*[1] nace de la alianza entre la empresa *Elsevier B.V.*[2] y el grupo de investigación *Scimago*[3]. Como resultado de esta sinergia se ha desarrollado una plataforma de indicadores científicos a partir de la información contenida en la base de datos *Scopus*[4].

La plataforma toma su nombre del indicador *Scimago journal rank* (*SJR*), elaborado por el grupo a partir del algoritmo *Page rank*, que representa la visibilidad de las revistas contenidas en la base de datos desde 1999 hasta 2006.

Alternativa al *JCR*

En la página principal del portal se encuentran dos opciones principales: el ranking de revistas y el de países. Como se puede suponer el portal *SJR* nace con la vocación de representar una alternativa *open access* a los productos de *Thomson Scientific JCR* y *ESI*[5].

El ranking de revistas se puede apreciar en la figura 1 (*Journal indicators*). En él aparecen unos menús desplegables tipo combo a través de los cuales se puede filtrar el ranking por una de las grandes áreas temáticas (27), una categoría temática (295), el país y el año. Hay un combo adicional que permite el ordenamiento del ranking bajo diferentes campos: *SJR*, citas por documento, índice h, título, documentos, documentos citables, y total de citas. También existe la posibilidad de recortar el ranking estableciendo un valor mínimo de umbral.

La apariencia del segundo de los *rankings* se puede apreciar en la figura 2 ("*Country Indicators*"). Los filtros son similares a los del caso anterior, aunque los campos mostrados son diferentes. Acá se incluyen, además del nombre de cada país, su producción total en número de documentos, la cantidad de citas recibidas, las citas por documentos, y el índice h.

Por otro lado, hay que destacar la opción "*Journal Search*", que permite buscar por una revista en particular, tanto por su título como por su ISSN. El resultado consiste en un resumen de todos los indicadores de esa revista en todo el período de cobertura, así como las representaciones gráficas de algunos de los mismos. Este es un breve informe al que también se puede acceder haciendo click en cada uno de los títulos del *ranking* de revistas.

En este informe se encuentran los indicadores ya presentados en

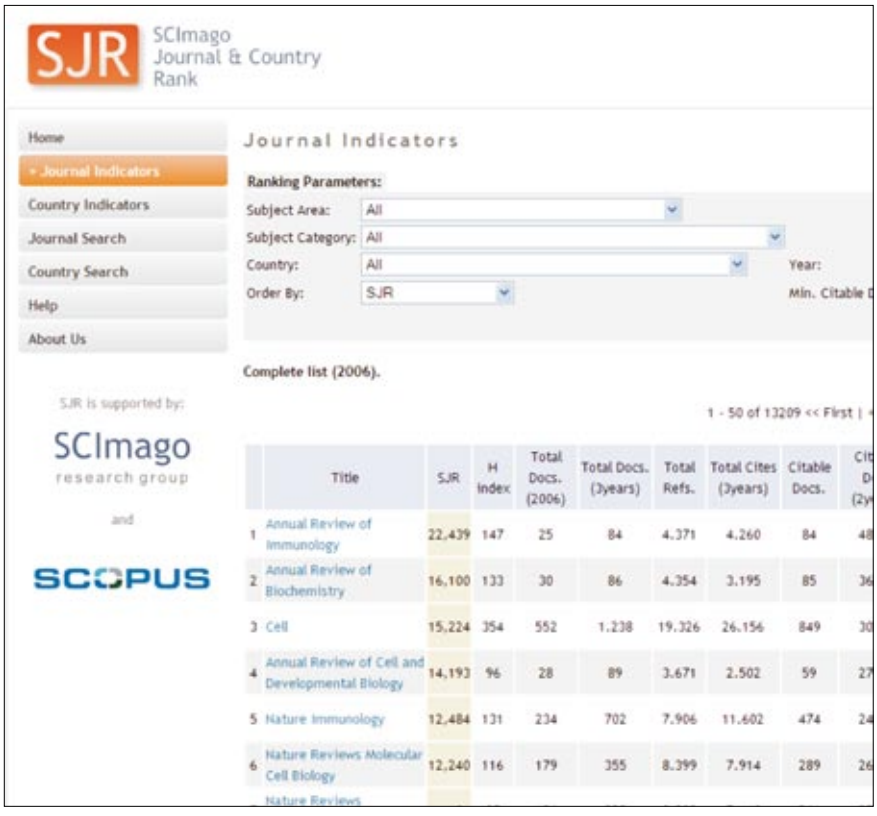


Figura 1. Ranking de revistas

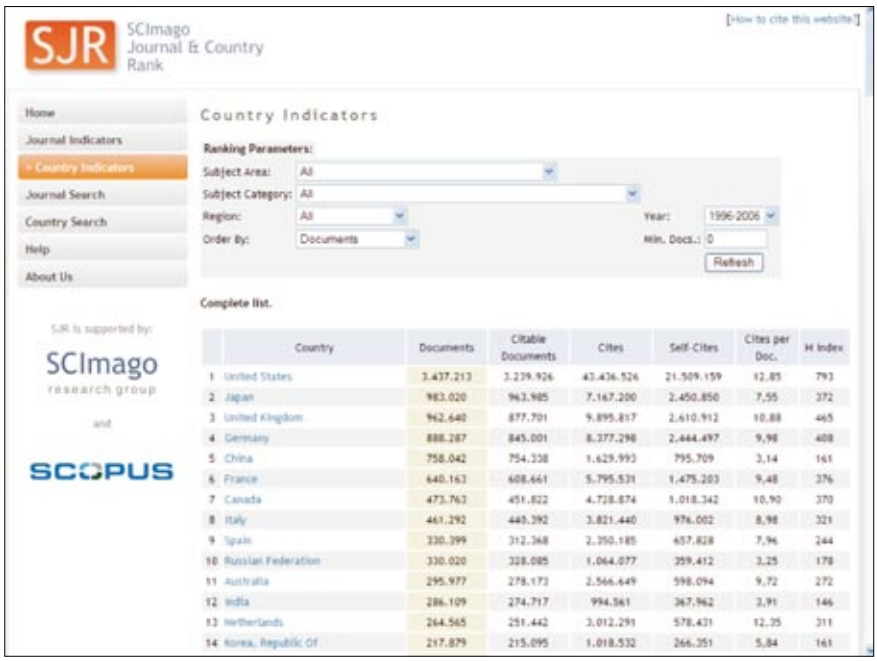


Figura 2. Ranking de países

el ranking principal junto a otros diferentes, discriminados por años. Entre ellos encontramos:

– *SJR*: es un indicador que expresa el número de enlaces que una revista recibe a través de la citación ponderada de sus documentos en relación con el número de documentos publicados en el año por cada publicación. La ponderación de las citas se hace en función de las que recibe la publicación citante.

– *Journal documents*: Número de documentos publicados. Para el cálculo de este indicador se tienen en cuenta todos los tipos de documentos incluidos en cada ejemplar de la revista en el año seleccionado.

– *Journal documents (3years)*: Número de documentos publicados en 3 años. Para el cálculo de este indicador se tiene en cuenta el acumulado de los trabajos (cualquier tipo de documento) publicados en los tres años anteriores.

– *Journal references*: Número total de referencias. Para el cálculo de este indicador se realiza un sumatorio del total de cualquier tipo de referencia bibliográfica incluida en los documentos de la revista en el año seleccionado.

– *Total cites (3years)*: Total de citas realizadas desde la revista en 3 años. Para el cálculo de este indicador se tienen en cuenta los documentos de cualquier tipo aparecidos en el año seleccionado en una publicación y las referencias bibliográficas que han realizado a cualquier documento publicado en los tres años anteriores.

– *Autocitas*: Número de citas recibidas por la publicación cada año, que proceden de la misma publicación.

– *Citable documents*: Total de artículos y reviews publicados en tres años. Para el cálculo de este indicador se han tenido en cuenta exclusivamente dos tipologías documentales: los artículos científicos y las revisiones de los tres años anteriores al año seleccionado.

– *Cites x document (4years)*: Promedio de citas por documento en cuatro años. Para el cálculo de este indicador se han tenido en cuenta el número de citas recibidas en los cuatro años anteriores y el número de documentos publicados en el año seleccionado.

– *Cites x document (3years)*: Promedio de citas por documento en tres años. Para el cálculo de este indicador se han tenido en cuenta el número de citas recibidas en los tres años anteriores y el número de documentos publicados en el año seleccionado.

– *Cites x document (2years)*: Promedio de citas por documento en dos años. Para el cálculo de este indicador se han tenido en cuenta el número de citas recibidas en los dos años anteriores y el número de documentos publicados en el año seleccionado

– *References x document*: Promedio de referencias por documento. Para el cálculo de este indicador se divide el total de referencias incluidas en los documentos de la revista en el año seleccionado entre el total de documentos publicados en ese mismo año.

Por lo que afecta a la función *Country search*, despliega un conjunto de indicadores científicos absolutos y relativos de cada país representado en la base de datos, con los correspondientes gráficos que facilitan su interpretación.

Como puede apreciarse, el portal presenta una amplia gama de indicadores. Varios de ellos son variaciones en torno al impacto utilizando diferentes ventanas de citación, pero el punto fuerte es el *Scimago journal rank* (*SJR*), que se basa en el algoritmo de *Page rank* para ponderar las citas que forman parte del cálculo. Sobre este indicador entraremos en detalle en el próximo artículo.

Notas

1. http://www.scimagojr.com/
2. http://www.elsevier.com/
3. http://www.scimago.es/
4. http://www.scopus.com/
5. http//www.isiknowledge.com/

Grupo Scimago (Imago scientiae o visualización de la ciencia)
scimago@ugr.es
http://www.scimago.es
http://www.atlasofscience.net

Gestores personales de bases de datos de referencias bibliográficas: características y estudio comparativo

Por **Emilio Duarte-García**

Resumen: *Se estudian las características comunes y específicas de los gestores personales de bases de datos de referencias bibliográficas más utilizados: Reference Manager, EndNote, ProCite, RefWorks y EndNote Web. Los apartados analizados son: la entrada de datos, el control de autoridades, los comandos de edición global, la personalización de algunos aspectos de las bases de datos, la exportación de las referencias, la visualización de los registros, la inserción de citas bibliográficas y la generación automática de bibliografías.*

Palabras clave: *Reference Manager, EndNote, ProCite, RefWorks, EndNote Web, Bases de datos bibliográficas, Gestores de referencias bibliográficas, Recuperación de la información, Formatos bibliográficos.*

Title: Personal managers of bibliographic reference data bases: Characteristics and comparative analysis

Abstract*: Both the shared and unique characteristics of the most widely used personal managers of bibliographic reference data bases are analysed: Reference Manager, EndNote, ProCite, RefWorks and EndNote Web. The aspects considered include: data input, authority control, global editing commands, personal configuration of the data bases, reference exportation, visualization of records, insertion of bibliographic references and automatic generation of bibliographies.*

Keywords*: Reference Manager, EndNote, ProCite, RefWorks, EndNote Web, Bibliographic data bases, Bibliographic reference managers, Information retrieval, Document format.*

Duarte García, Emilio. "Gestores personales de bases de datos de referencias bibliográficas: características y estudio comparativo". En: *El profesional de la información,* 2007, noviembre-diciembre, v. 16, n. 6, pp. 647-656.

DOI: 10.3145/epi.2007.nov.12

***Emilio Duarte García**, licenciado en geografía e historia (1989) y master en biblioteconomía y documentación (1990-1992), ambos por la Universidad del País Vasco, y diplomado en biblioteconomía y documentación por la Universidad de Salamanca (1999). Ha trabajado en bibliotecas públicas y en la Biblioteca de la Universidad del País Vasco, donde actualmente participa en la formación de usuarios en los programas Reference Manager y EndNote Web.*

1. Introducción

En los últimos años se ha generalizado la utilización de los denominados gestores personales de bases de datos de referencias bibliográficas, que son programas que permiten a investigadores, especialistas y profesionales almacenar las referencias bibliográficas obtenidas durante la búsqueda documental (**Armeteros**, 2004) para su posterior gestión y manipulación, con la finalidad de insertar citas y elaborar bibliografías de acuerdo con los formatos de descripción que exigen las diferentes revistas científicas.

El objetivo de este estudio ha sido analizar las características, tanto comunes como específicas, de los más empleados. Por motivos metodológicos se detallan en primer lugar las propiedades que comparten todos los programas en cada uno de los apartados que se han contemplado, y seguidamente se señalan las particularidades que los distinguen entre sí.

Los programas analizados, teniendo en cuenta sus versiones, han sido[1]: *ProCite 5*, *Reference Manager 11*, *EndNote X*, *RefWorks 4.2* y *EndNote Web 1.3*. Los tres primeros necesitan ser instalados en el ordenador para poder ser utilizados y únicamente *EndNote* dispone de versión tanto para *Windows* como para *Mac*, el resto sólo para *Windows*. Por su parte, *RefWorks* y *EndNote Web* se consultan vía internet, permitiendo ambos el uso de *Windows* y *Mac*.

Artículo recibido el 15-05-07
Aceptación definitiva: 25-06-07

2. Funciones y características de los gestores de bases de datos de referencias bibliográficas

Antes de seguir es obligado señalar que en todas estas aplicaciones se pueden observar dos partes claramente diferenciadas, pero a su vez íntimamente ligadas. Por un lado, el gestor de bases de datos documentales, que administrará las referencias bibliográficas que vayamos añadiendo y, por otro, una herramienta que incorporada en nuestro procesador de textos, permitirá la inserción de citas y la generación de bibliografías de forma automática y relativamente sencilla en función de los datos introducidos en nuestra base.

"Son programas gestores de bases de datos documentales que además permiten insertar citas y generar bibliografías de forma automática"

2.1. Entrada de datos

Suele efectuarse por medio de registros con estructuras predefinidas para los diferentes tipos de publicaciones: artículos de revistas, monografías, capítulos de libros, tesis, informes, etc. Cada uno de ellos cuenta con campos comunes, como los de autor, año, fuente, resumen y palabras clave, junto a otros específicos en función del tipo de publicación (**Bravo**, 2007).

La entrada de datos puede realizarse de dos formas: manual y automática. Si se hace manualmente, el primer paso consistirá en rellenar el campo que determina el tipo de documento que vamos a describir; de esta manera quedarán automáticamente seleccionados aquellos campos que estén asociados a dicho tipo documental. El hecho de que los registros dispongan de esas estructuras predefinidas facilita significativamente esta tarea, y además todos los programas proporcionan un número considerable de plantillas diferentes que permiten introducir desde las referencias de los documentos más comunes, como libros, artículos de revista, actas de congresos, etc., hasta las de materiales menos estándar, como programas de ordenador, audiovisuales, mapas, partituras, etc.

Si se hace un análisis individualizado:

– *Reference Manager* distingue hasta 35 tipos documentales y suma 37 campos.

– *EndNote*: 41 tipos documentales y 52 campos.

– *ProCite*: 39 tipos documentales, con la posibilidad de crear plantillas para tipos nuevos, y 45 campos.

"La introducción de los datos puede hacerse de forma manual y también automáticamente desde fuentes de información externas"

– *EndNote Web*: 39 tipos documentales y 50 campos.

– *RefWorks*: 31 tipos documentales y 54 campos.

Tanto el número de referencias como la longitud de los campos es ilimitado, salvo en el caso de *EndNote Web*, que establece un límite de 10.000 referencias y una capacidad máxima de 64.000 bytes por campo. De cualquier manera, ese tamaño suele ser más que suficiente para la introducción de los datos necesarios.

Entre los campos que habría que completar en estas bases de datos hay algunos que deben destacarse por su interés: aquellos en los que es factible la creación de enlaces a direcciones url, archivos pdf, documentos completos y archivos de imágenes. De esta forma, y desde el mismo registro bibliográfico podríamos acceder, por ejemplo, al texto completo del documento referenciado.

Por otro lado, un aspecto muy interesante para los usuarios es la posibilidad de introducir referencias bibliográficas en sus bases de datos de forma automática. Existen dos opciones:

1. De bases de datos online a través de la interfaz que los propios programas han diseñado, y que automáticamente almacenará la información necesaria para establecer las conexiones, realizará las búsquedas e importará las referencias a nuestra base de datos.

Todos los programas tienen conexión a *PubMed*, a *Medline*, a catálogos que utilicen el protocolo de comunicación estándar internacional Z39.50 y, salvo *RefWorks* y *ProCite*, a las bases de datos del *ISI Web of Knowledge* (*WOK*)[2]. En todo caso, es interesante observar cómo está estructurado su acceso: mientras que en *EndNote*, *RefWorks* y *EndNote Web* aparece el listado alfabéticamente ordenado de todas las bases disponibles, en *ProCite* y *Reference Manager* hay apartados de búsqueda diferentes para buscar por un lado en *PubMed*, por otro en catálogos Z39.50, y aunque sólo en el caso de *Reference Manager*, en el *WOK*.

En cuanto a la conexión a catálogos Z39.50, únicamente *ProCite* y *Reference Manager* permiten realizar búsquedas en varias bases de datos de manera simultánea, mientras que el resto sólo soportan la consulta a una única en cada ocasión.

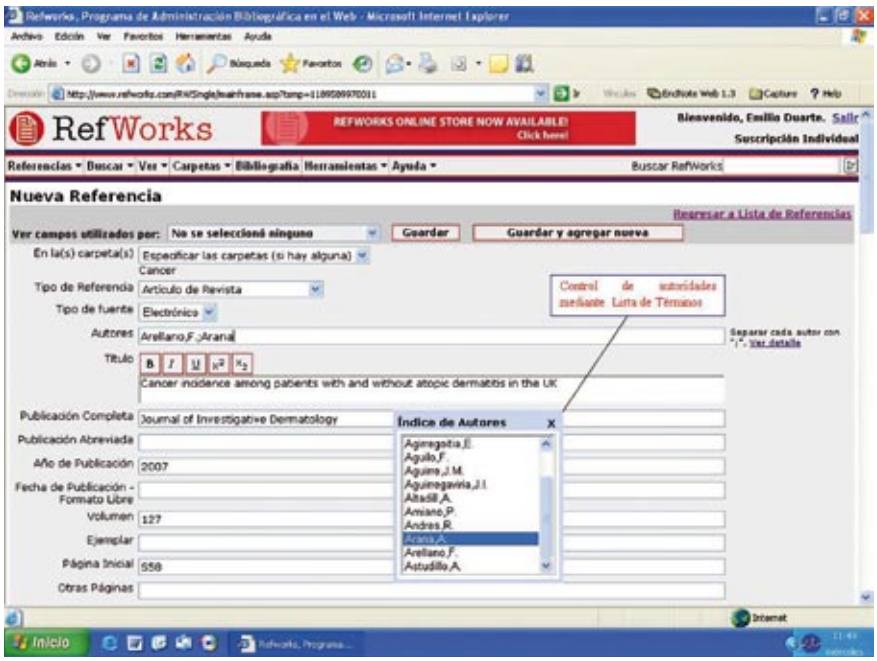


Figura 1. Registro utilizado para la entrada manual de datos y Lista de términos para el control de autoridades en RefWorks

EndNote y *EndNote Web* son los únicos gestores que ofrecen la posibilidad de confeccionar listas de favoritos, es decir, un listado con las bases de datos que más utilizamos. En cambio, en todos se pueden crear nuevas conexiones a catálogos Z39.50, bien realizándolas uno mismo o solicitándoselas al proveedor del programa.

Cuando obtenemos referencias a través de esta función, los programas las adaptan de manera automática al formato que cada uno tiene definido por defecto y construyen una base de datos temporal con las referencias descargadas para que las exportemos o copiemos a cualquiera de las nuestras (**Codina**, 2000).

RefWorks permite también la importación de canales RSS[3] (*really simple sindication*) ya que lleva integrado un lector de RSS. Algunos proveedores de información posibilitan la creación de *feeds* específicas según un criterio de búsqueda. Algunos ejemplos serían: *NLM Pubmed*, *Scopus*, *Nature*, etc. (*RefWorks*).

2. Desde fuentes de información externas, es decir, desde bases de datos, revistas electrónicas, etc., a las que nos conectaremos online sin utilizar el interfaz del gestor de referencias bibliográficas. En este caso se ofrecen también dos opciones:

a. Existen muchas fuentes (*Web of Science*, *WebSpirs*, *Ebsco Online*, catálogos de bibliotecas, etc.) que facilitan la exportación directa a este tipo de programas desde su propia interfaz de usuario mediante comandos igual o similares a "Export to reference software". En estos casos, una vez seleccionados los registros, y mediante el comando antes citado, la exportación se realizará a aquel programa que hayamos escogido.

b. Para los recursos que no tengan la función de exportación directa, deberemos descargarnos las citas en un archivo como fichero de texto con un formato estructurado (extensiones .txt o *.ris*, por ejemplo). Posteriormente habrá que importarlo a nuestra base de datos personal. En este caso, todos los programas van a funcionar de forma muy similar. En primer lugar se seleccionará el archivo de texto, y seguidamente la base de datos concreta a donde queremos importarlo; finalmente, se elegirá el filtro de importación correspondiente.

Las bases de datos bibliográficas presentan los datos estructurados de formas diversas y, por tanto, será necesario contar con un filtro apropiado que reconozca la estructura de datos de la referencia que se va a importar. Así, un filtro de importación permite extraer información de dichos archivos de texto e introducirlos en los campos pertinentes de nuestra base de datos personal. Los campos del filtro deben corresponderse

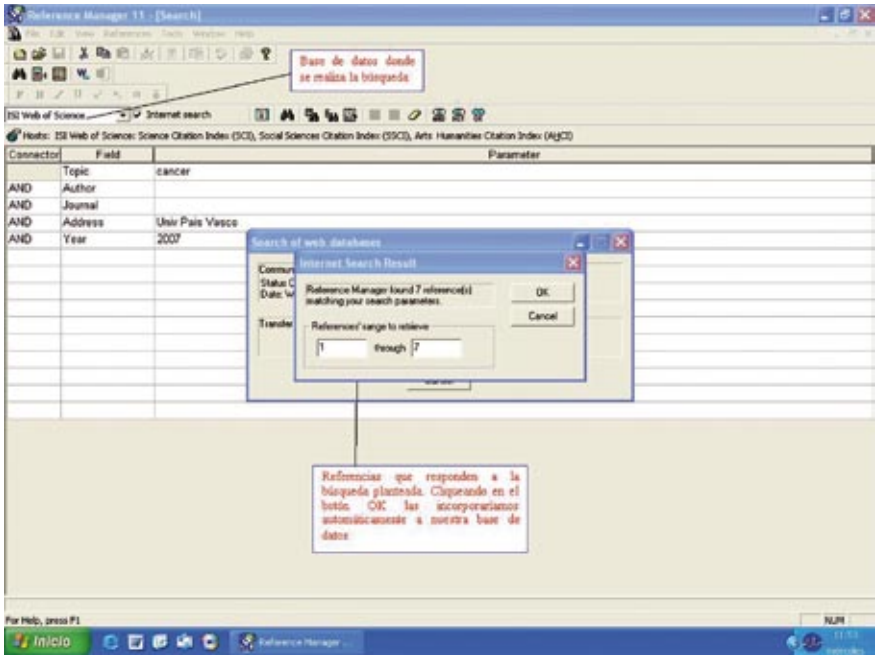


Figura 2. Búsqueda de registros en las bases de datos del ISI Web of Science desde el interfaz de Reference Manager

"Se asegura la coherencia en la entrada de datos mediante el control de autoridades ejercido desde las listas de términos"

con los del archivo que se pretende exportar. El nombre de cada filtro suele indicar el servicio para el que está configurado. Todos los programas presentan un número considerable de filtros confeccionados previamente y, si no encontramos el apropiado, podemos optar por crear uno propio (caso de *Reference Manager*, *ProCite* y *EndNote*) o pedir al suministrador del programa que lo cree (*EndNote Web* y *RefWorks*).

EndNote Web tiene habilitada, únicamente para Pc, la opción del botón "Capture", que se encuentra en la barra de herramientas que el programa muestra en nuestro navegador, y permite capturar fácilmente la información bibliográfica del registro existente en la página web que tengamos activa. Su funcionamiento y características son los siguientes:

1. Al realizar la captura debemos tener a la vista el registro completo, y solamente salva uno cada vez, por lo que no se pueden salvar referencias marcadas de la página de resultados.

2. Si hacemos clic en el citado botón nos creará una referencia cuyos campos completaremos o modificaremos. Finalmente, confirmaremos que el tipo de referencia aparecido por defecto es el correcto y que los campos obligatorios marcados con un asterisco han sido completados.

2.2. Control de autoridades

Aunque este punto se encuentra relacionado con la entrada de datos, se ha preferido diferenciarlo en un nuevo apartado para analizarlo por separado en cada uno de los programas.

Únicamente *EndNote Web* carece de esta función; los demás disponen de lo que se denominan listas de términos, que son listados vinculados a los contenidos de los campos de palabras clave, autor y revista (en el caso de *ProCite* también a los de título y tipo documental) que permiten almacenar palabras clave, nombres de autor, revista, título y tipo de documento. Estas listas se gestionan de forma automática: cuando se crea una nueva base de datos están vacías, y conforme se introducen las referencias se van añadiendo automáticamente los nuevos términos.

Su utilidad es clara: introducir consistentemente las referencias en nuestra base de datos. La forma de hacerlo varía según los programas; así, mientras en *Reference Manager* y *RefWorks* nos saldrá automáticamente el listado de los términos ya introducidos en cuanto tecleemos algún carácter, en *EndNote* el programa intentará completar el texto sugiriendo la palabra más cercana de la lista vinculada a ese campo, de tal forma que si queremos ver el resto de los términos deberemos movernos con las teclas de navegación. Finalmente, en *ProCite* no nos saldrá por defecto el listado de los términos, por lo que se hará necesario utilizar los botones situados junto al campo correspondiente para poder acceder a ellos. En todos los casos si el término ya existe se utiliza, si no, se crea nuevo, y al guardar la referencia se añadirá automáticamente a la lista.

2.3. Búsquedas en las bases de datos personales

Estas consultas son similares en todos los programas analizados. Se efectúan en pantallas preconfiguradas y se puede elegir por un lado entre realizarlas en una base de datos o en varias a la vez, y por otro en toda la base de datos o en campos determinados. También es posible construir búsquedas complejas que involucren varios campos y términos utilizando los operadores booleanos clásicos ("and", "or", "not"), relacionales, truncamientos y máscaras o comodines, además de guardar las estrategias de búsqueda para su uso posterior.

Reference Manager, *ProCite* y *EndNote* permiten crear un índice de referencias dentro de una base de datos para guardar un conjunto de ellas, facilitando de esta forma una rápida localización posterior del grupo o la ejecución de determinadas operaciones sobre el mismo. Por ejemplo, cuando una entidad trabaja con dos aspectos o disciplinas, podemos incluirlos en la misma base de datos y a la vez introducirlos en un grupo para operar con ellos separadamente (**Alonso**, 2006).

Por último, en todos los programas (excepto *EndNote Web*) es posible acceder a la lista de términos de los campos indizados, pero sólo *Reference Manager* ofrece la opción de integrar en la búsqueda más de un término de la lista unido por los operadores "and" y "or". También es *Reference Manager* el único que permite establecer relaciones de sinonimia entre términos de una misma lengua que posean idéntico significado, o entre un vocablo y sus equivalentes en otros idiomas, así como vincular distintas siglas con su expresión desarrollada. Cuando se combinan dos términos con referencias cruzadas, ambos aparecen en el listado y cada uno de ellos incluye al otro en la lista de sinónimos. Esta operación puede ser ejecutada de forma manual o automática.

2.4. Detección de duplicados

Cuando se introducen referencias, tanto manualmente como importándolas automáticamente de varias fuentes, es muy fácil acabar insertando registros repeti-

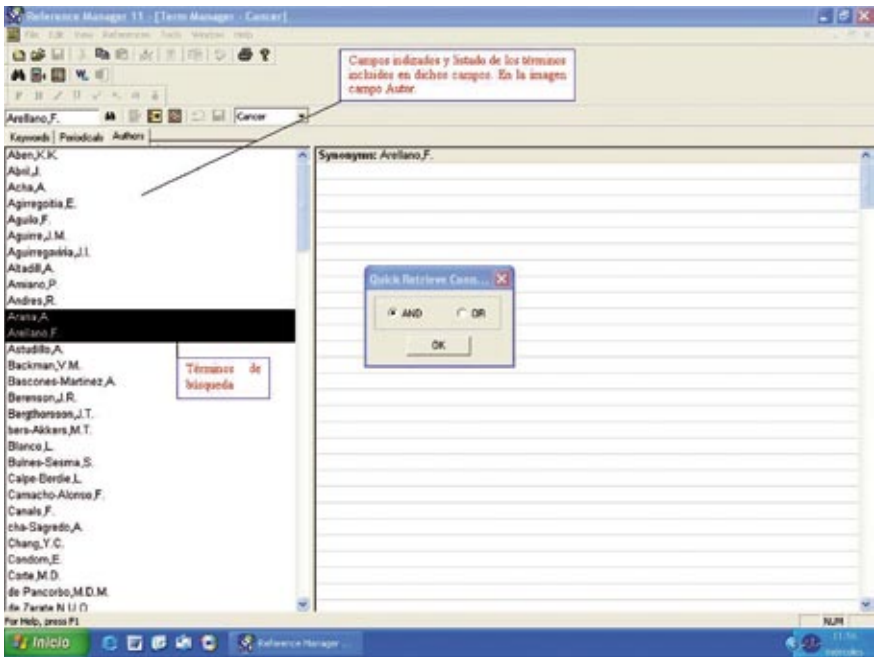


Figura 3. Búsqueda en la Lista de Términos de los campos indizados utilizando los operadores booleanos And y Or en Reference Manager

dos. Para evitarlo, los programas disponen de una serie de mecanismos con los que detectar los duplicados. En primer lugar, definen por defecto lo que consideran que son, es decir, aquellas referencias que pertenezcan al mismo tipo documental (ej.: artículo de revista, libro) y en las que además coincidan el autor, título y año. Estos criterios, salvo en *EndNote Web* y en *RefWorks*, se pueden modificar.

La detección puede realizarse cada vez que introducimos manualmente un documento nuevo o lo editamos (es el caso de *Reference Manager* y *EndNote Web*). También cuando se copia algún registro resultante de una búsqueda online a través de la interfaz de los propios programas en una de nuestras carpetas o bases de datos (en todos salvo en *EndNote* y *ProCite*).

Una tercera manera de localizar duplicados es durante el proceso de importación de archivos de texto. Todos los programas salvo *ProCite* implementan esta función, pero unos con más opciones que otros. Por ejemplo, *Reference Manager* y *EndNote* pueden decidir que ante un duplicado se actúe de tres formas diferentes: que se importen todas las referencias, incluidas las duplicadas, lo que requerirá una revisión manual posterior; que se importen todas salvo las duplicadas (única opción existente en *EndNote Web* y *RefWorks*), o que las referencias duplicadas sean trasladadas a una nueva base de datos temporal, mientras el resto son importadas normalmente.

Finalmente, si a pesar de las medidas adoptadas para evitar la entrada de duplicados éstas han sido superadas tendremos la opción, en la totalidad de los programas, de buscarlos.

Reference Manager permite además la combinación de referencias duplicadas, es decir, dos referencias referidas a un mismo documento que contengan distinta información y que interese unir en una sola que reúna tanto los campos y datos comunes como los que difieren de una referencia a la otra (*Manual de Reference Manager v. 11*).

2.5. Comandos de edición global

Salvo *EndNote Web*, todos permiten el uso de comandos u órdenes que actúan de forma general en toda la base o bases de datos para: buscar y reemplazar texto en múltiples referencias (única posibilidad existente en *RefWorks*), copiar campos entre referencias o mover información de un campo a otro en múltiples referencias.

2.6. Visualización de los registros

Tanto en *Reference Manager* como en *EndNote* la ventana de visualización de referencias aparece dividida en dos partes (en el caso de *ProCite* hay que indicárselo expresamente). Uno de los paneles muestra la lista de referencias, es decir, un listado en el que cada línea corresponde a una referencia distinta, con la información de los campos ordenada en columnas. Los campos que aparecen por defecto varían según los gestores: *Reference Manager* y *ProCite* usan tres (el primero los de número de identificación de la referencia, autor primario y título, y el segundo los de autor primario, título y fecha), mientras que *EndNote* utiliza seis (autor primario, año, título, revista, tipo de referencia y url). Este panel se puede personalizar, y si se clica sobre cualquiera de las cabeceras de las columnas se puede ordenar alfabética o numéricamente la lista, bien en sentido ascendente o descendente. El otro panel presenta el contenido completo de la referencia y también se puede personalizar.

"La detección de duplicados, las diferentes posibilidades existentes en la búsqueda y visualización de los registros y la opción de utilizar comandos de edición global son herramientas interesantes"

EndNote Web únicamente maneja una ventana de visualización, que coincidiría, por como está dispuesta la información, con la lista de referencias de los anteriores. Muestra cuatro columnas por defecto: autor primario, año, carpeta en la que está la referencia y título. Utiliza las cabeceras de las columnas para poder cambiar el criterio de ordenación de la lista.

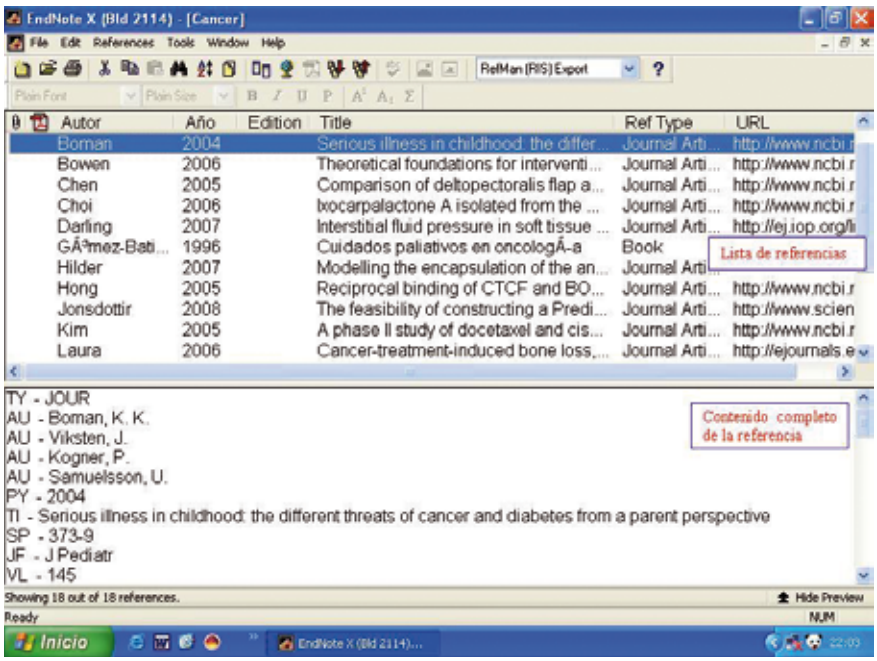


Figura 4. Ventana de visualización de registros en EndNote

Por último, *RefWorks* ofrece tres tipos de vista: "Una lista/vista de cita", que dispone en una sola línea los datos de los campos de autor, fecha y título; "Vista uniforme", con la información de los campos de título, autor y fuente, e indicando el nombre de la base de datos en la que se encuentra; y "Vista completa", que presenta el contenido completo de la referencia. En las dos últimas, cuando se visualice una referencia que ha sido importada desde *PubMed* aparecerán los enlaces a dicha base de datos.

En todos los casos se pueden ordenar los registros según diferentes criterios: autor primario, año de publicación (descendente o ascendente), tipo de referencia, etc.

2.7. Personalización de aspectos de la base de datos

Todos los programas permiten personalizar utilidades relacionadas con la visualización de los resultados en pantalla, pero únicamente *EndNote*, *ProCite* y *Reference Manager* confieren la posibilidad de modificar aspectos que suponen un valor añadido para dichos programas.

Se puede configurar la lista de referencias indicando qué campos queremos que aparezcan, y cambiando el orden y el nombre de las cabeceras. También es posible modificar la información sobre los distintos tipos de referencia (salvo en el tipo *Generic*), seleccionando los campos que se desea que aparezcan en cada modelo de registro, así como renombrar las etiquetas de los campos (**Codina**, 2000)**.** Finalmente, sólo con *Reference Manager* se puede cambiar el orden de aparición de los campos y ocultar los tipos de referencia que no se usen, acortando la lista, facilitando y estandarizando la entrada de datos y renombrando los tipos de referencia.

Por último, *Reference Manager* personaliza también la lista de campos y su ordenación en la columna *Field*, en la ventana de búsqueda de referencias (*Search*).

2.8. Exportar y compartir referencias bibliográficas

Son otras dos funciones de enorme interés.

1. La opción de exportar permite copiar las referencias de una base de datos para crear un archivo de texto en función del formato de salida seleccionado. Son varios los formatos que cada programa tiene disponibles, siendo los coincidentes: *Medline*, formato de archivos de texto que exporta campos de información a la *National Library of Medicine*; *Tab delimited***,** que separa los campos de información mediante tabulaciones; y *RIS*, que utiliza *Reference Manager*.

"Permiten exportar y/o compartir referencias fácilmente con otros investigadores"

Además de los citados, cada programa cuenta con su propio repertorio: *Reference Manager* tiene *Comma Delimited*, que permite separar los campos mediante comas, y xml que exporta referencias desde *Reference Manager* a una gran cantidad de aplicaciones en internet. *EndNote Web* dispone de *BibTex*, una herramienta para dar formato a las listas de referencias usadas por el sistema de preparación de documentos *LaTex*, que es un lenguaje de programación de excelente calidad orientado a la escritura de textos (**Cascales**, 2006), y de *Refer Export*, formato de exportación de *EndNote*.

RefWorks cuenta con *BibTex*, xml *y Citation List*, que crea una lista que incluye la información de los campos de número de identificación del documento, autor primario, título y año de publicación. Por último, *EndNote* dispone de *BibTex*, xml, html, *Text File* (.txt) y *rich text format* (.rtf), formato de intercambio que mantiene los estilos (como la negrita, la cursiva, etc.).

2. La opción de compartir las bases de datos o carpetas de referencias bibliográficas creadas por los usuarios únicamente está implementada en *EndNote Web* y *RefWorks*.

Una vez se ha decidido la/s carpeta/s o base de datos que se quiere compartir y los usuarios a los que se concederá tal permiso, hay que tener en cuenta que el acceso a ellas permite únicamente su lectura, es decir, que no se podrán modificar los registros, pero sí buscar y visualizar el contenido de las referencias, generar e imprimir una bibliografía, exportarlas, copiarlas a una carpeta o base de datos propia, etc.

La función de compartir es muy interesante, ya que nos va a permitir, por ejemplo: colgar bibliografías temáticas en una página web; proporcionar acceso rápido a la información para investigadores de diferentes instituciones que colaboren en un mismo proyecto; facilitar un enlace a la bibliografía publicada por un departamento o investigador en concreto; publicar una base de datos interna de referencias dentro de una misma organización, etc. Por tanto, los beneficios resultantes son muy variados: favorece la comunicación y difusión de la información; aporta un entorno de investigación cooperativo accesible desde cualquier lugar a través de internet; permite el intercambio de información entre bases de datos de una manera rápida y fácil; ofrece un punto de acceso centralizado a fuentes de información necesarias puntualmente sobre una asignatura o un proyecto de investigación concreto, etc. (*RefWorks*).

2.9. Las citas bibliográficas

Una vez instalados los programas, en el procesador de texto *Microsoft Word* aparecerá una barra de herramientas con una serie de comandos que nos permitirán insertar las citas y elaborar las bibliografías a partir de las referencias incluidas en las bases de datos (*Manual de Reference Manager v. 11*). A las utilidades de la barra de herramientas también se puede acceder mediante el menú de herramientas del procesador de textos.

Esta utilidad, denominada "Cite while you write" o "Write-n-cite" dependiendo de los programas, de lo que se encarga en definitiva es de añadir códigos (que quedan ocultos) de campo a las citas. Su función es permitir al programa dar y quitar formato al texto y reformatear las citas dentro del procesador de textos.

Las citas pueden configurarse de dos maneras:

1. Formato temporal: contienen la información necesaria para poder localizar una única referencia de la base de datos y determinar cuáles se han de incluir en la bibliografía. Suele constar de los datos que identifican al autor (normalmente el apellido del primer autor), el año de publicación y un identificador de referencia, que es el número que esa referencia bibliográfica tiene en la base de datos; y todo ello entre delimitadores, normalmente unas llaves. Por ejemplo: *{Barlow, 1990 11/id}*[4].

2. Citas formateadas: incluyen códigos de campo ocultos para poder generar la bibliografía, añadir más citas o modificar su estilo de salida. Pueden aparecer en un estilo autor/año -*(Barlow, 1990)*— o en estilo numérico –*[1]*–. Su aspecto dependerá del estilo bibliográfico de salida seleccionado.

El estilo de salida es un formato que se le da a las citas y a las bibliografías, y que responde a las normas del editor o revista que lo haya definido (por ej.: "Nature style" pertenece a la revista *Nature*). En todos los programas hay numerosos estilos bibliográficos para elegir, y están identificados con el nombre del creador (por ej.: "Chicago style" corresponde al *Chicago manual of style*). Si entre los preexistentes no encontráramos el que necesitamos, podríamos crear nosotros mismos otros nuevos o solicitárselos al proveedor del programa.

El hecho de que las referencias aparezcan con un formato o con otro dependerá de que hayamos seleccionado o no por defecto la opción de formateo automático a la hora de insertarlas. Si la deshabilitamos, todas las citas llevarán un formato temporal, no se generará de forma automática una bibliografía al final del documento y será necesario formatearlo. En cambio, si la activamos, cada vez que introduzcamos una cita quedará ya formateada y la bibliografía se listará automáticamente al final del documento. Por lo tanto, antes de insertar referencias, habrá que decidir la forma en que queremos hacerlo.

Las citas se pueden incluir en cualquier parte del documento, sea en el cuerpo del mismo o en una nota al pie de página (salvo en *EndNote Web*). Para ello basta con posicionar el cursor en el lugar deseado y seleccionar la opción correspondiente en la barra de herramientas. Se puede insertar una cita única: *(Córdoba González, 2001)*; o bien múltiples referencias en una: *(Gazpio and Álvarez, 2002; Gómez Hernández and Benito Morales, 2003; Hernández Salazar, 2003)*. Cuando se dé el caso de una cita múltiple, ésta se organizará según los requerimientos del estilo de salida establecido.

Las referencias introducidas en un documento pueden cambiarse en cualquier momento. Existe la posibi-

lidad de modificar, añadir, ordenar y eliminar referencias de una cita múltiple.

En el caso de *RefWorks* hay que señalar que cuenta con un sistema de inserción de citas mucho más lento y complicado que el resto de programas. En primer lugar, es preciso efectuar un número mayor de pasos para poder hacerlo, y además estará siempre en formato temporal y será necesario además grabar el documento de *Word* donde insertamos las citas para después poder generar la bibliografía. Una vez originada, se crea automáticamente un nuevo documento con las referencias formateadas y la bibliografía al final del documento. Si se quiere editar las citas debe realizarse en el documento original y posteriormente formatearlo de nuevo.

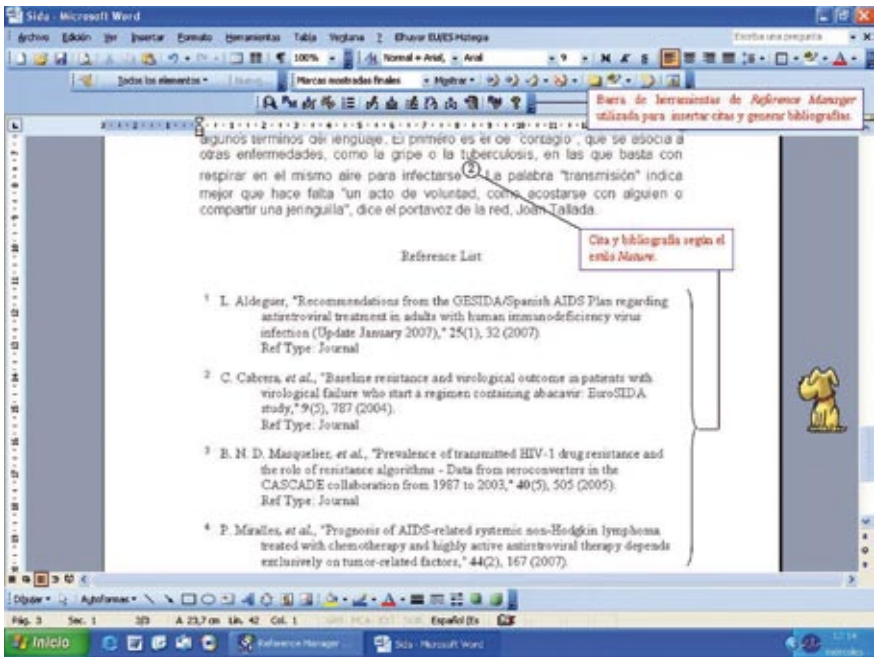


Figura 5. Citas formateadas insertadas en el texto, junto a la bibliografía generada de forma automática. Estilo de salida, Nature. Programa utilizado, Reference Manager

Tanto *EndNote* como *RefWorks* y *EndNote Web* contemplan la posibilidad de insertar citas y generar bibliografías en procesadores de texto diferentes a *Microsoft Word*. En todos los casos deben incluirse manualmente en formato temporal y posteriormente el documento se formatea para que las referencias aparezcan también formateadas y la bibliografía quede listada al final del documento.

Todos los programas permiten también:

– Revertir las citas formateadas, es decir, cambiar una ya formateada por su correspondiente temporal y eliminar la bibliografía. Esta opción puede ser interesante cuando se quiere confeccionar una bibliografía general al final de un documento que ha sido redactado por varios autores, y en el que cada uno de ellos ha elaborado una parte con sus propias citas y su bibliografía. El procedimiento consistirá en pasar las citas a formato temporal, copiar y pegar los textos en el orden que corresponda y, finalmente, formatear de nuevo el documento para generar la bibliografía al final del mismo.

– Eliminar códigos de campo. Hay publicaciones que requieren documentos sin códigos de campos. Mediante esta herramienta se guarda una copia del documento con las citas sin formato y la bibliografía como texto.

– Crear una base de datos o una carpeta con las citas contenidas en un documento de *Word*.

2.10. Generar bibliografías

Una de las opciones más apreciadas en este tipo de programas, junto con la anterior de insertar citas de forma sencilla, es la de generar bibliografías a partir de bases de datos personales con los diferentes estilos de citación que utilizan habitualmente las revistas (**Armenteros**, 2006).

Se van a poder confeccionar de varias formas:

1. En función de las citas encontradas en el documento. Como ya hemos comentado en el apartado anterior, las formateadas producen al final del documento una bibliografía. Estas citas van acompañadas de códigos, y cada uno de ellos representa una referencia en la base de datos. El programa los remplaza automáticamente con números consecutivos (o autor/año), y añade al final del documento la lista de referencias formateadas y ordenadas adecuadamente, conforme aparecen en el texto o alfabéticamente, de acuerdo con el estilo seleccionado (**Armenteros**, 2006).

"Permiten insertar citas y generar una bibliografía tanto al final del documento como independiente, de una forma sencilla y automática"

Los estilos de salida determinarán qué información se incluye en la bibliografía, cómo se ordena y qué signos de puntuación y tipo y tamaño de letra se emplean.

2. Creación de una bibliografía independiente, es decir, aquella que no lleva asociado un documento con

citas insertadas a lo largo del texto. Así, una vez seleccionadas las referencias de nuestra base de datos, podremos generar listas en el estilo elegido. En este caso, y a diferencia de los documentos con citas insertadas, el estilo de salida deberá escogerse desde el propio programa de gestión de referencias bibliográficas (*Reference Manager v. 11*).

3. Creación de una bibliografía por materias (excepto *EndNote Web*), es decir, una lista de términos junto con las referencias asociadas a éstos. Se podrán generar basándose en cualquiera de los campos disponibles en la base de datos, si bien las bibliografías de materias típicas incluyen los contenidos de los campos autor, título de revista y palabras clave o materias.

4. Finalmente, pueden producirse listas de términos junto con el número de registros en los que aparecen.

2.11. Creación de manuscritos para editores con plantillas

Esta herramienta únicamente esta disponible en *EndNote* y proporciona numerosas plantillas que guían al usuario en la elaboración de documentos *Word* para ser entregados a los editores. Cada una de ellas recoge los requerimientos que cada editor exige para poder publicar en su revista.

Cuando se usan, muchos requerimientos editoriales ya están preestablecidos: márgenes, encabezamientos, paginación, espacios interlineados, el lugar donde se han de colocar las distintas partes del texto (título, resumen, agradecimientos, figuras y tablas, referencias bibliográficas), tipo de fuente y tamaño de letra (*EndNote*).

Se puede crear un documento *Word* utilizando estas plantillas, bien desde el mismo programa *EndNote*, mediante la ayuda de un asistente, o directamente desde el procesador *Word*, esta vez sin la ayuda de asistente alguno.

3. Conclusión

Si nos preguntáramos qué gestor de referencias es el más adecuado para una posible compra, tendríamos que valorar varias cuestiones:

– Todos los programas analizados son de calidad y ofrecen unas prestaciones que van a responder satisfactoriamente a las necesidades de los usuarios en cuanto a la creación y gestión de bases de datos personales, y a la posibilidad de insertar citas y generar bibliografías de manera sencilla y automática según el estilo bibliográfico seleccionado.

– Para usuarios de *Mac* las posibilidades se limitarían a tres: *EndNote*, *RefWorks* y *EndNote Web*.

Si hubiera que elaborar un ranking, a la cabeza estaría *Reference Manager*, a continuación *EndNote* y *RefWorks*, seguidos muy de cerca por *EndNote Web*, y finalmente *ProCite*.

– Todos se actualizan, y sirva como ejemplo que, salvo *ProCite*, ya pueden ser utilizados con *Windows Vista*.

– Además, es relativamente sencillo traspasar registros generados en uno de los programas a otro de la misma categoría, ya sea haciéndolo automáticamente desde la base de datos, o mediante los módulos de exportación e importación de registros.

Aunque posiblemente lo más objetivo sería utilizar las versiones de demostración[5] que todas las empresas tienen a su disposición vía internet, y de esta forma poder tomar una decisión con respecto a la compra de unos programas que, sin duda, son de gran utilidad, en mi opinión. Si hubiera que elaborar un ranking, a la cabeza del grupo estaría *Reference Manager*, a continuación *EndNote* y *RefWorks*, seguidos muy de cerca por *EndNote Web*, y finalmente, cerrando el pelotón *ProCite*.

Para finalizar, me gustaría expresar mi más sincero agradecimiento a **Laura García** por su traducción al inglés del resumen y las palabras clave, así como a **Lourdes Sáenz de Castillo**, por la lectura previa del presente artículo y sus certeros comentarios, así como a **Teresa Matellán** por su inestimable ayuda en la corrección estilística.

Notas

1. *Reference Manager*
http://www.referencemanager.com/
EndNote
http://www.endnote.com/
ProCite
http://www.procite.com/
RefWorks
https://www.refworks.com/
EndNote Web
http://www.endnoteweb.com/

2. Bases de datos referenciales con más de 8.500 revistas relacionadas con los ámbitos científicos de ciencia, tecnología y humanidades que pueden ser consultadas en:
http://www.accesowok.fecyt.es/login/

3. RSS es un formato basado en xml diseñado para la distribución de las noticias publicadas en sitios web, foros y weblogs que se actualicen con frecuencia. Un hilo o feed es un servicio gratuito que nos permite obtener automáticamente la última información de nuestras páginas web favoritas sin necesidad de abrir el navegador para comprobar si hay novedades (*IDG*).

4. Hay pequeñas variaciones, así, el aspecto que las citas en formato temporal tienen según el programa es: *Reference Manager*, *{Barlow, 1990 11/id}*;

EndNote, *{Barlow, 1990 #11}*; *EndNote Web*, *{Barlow, 1990}*; en *ProCite*, *{Barlow 1990 #11}* ó *{#11}*; y en *RefWorks*, *{{11 Barlow 1990}}*.

5. Las direcciones donde acceder a las versiones de demostración serían:
Reference Manager
http://www.referencemanager.com/rmdemo.asp
EndNote
http://www.endnote.com/endemo.asp
ProCite
http://www.procite.com/pcdemo.asp
RefWorks
https://www.refworks.com/RWSingle/newuser.asp?trial=y
EndNote Web
http://www.endnoteweb.com/enwtrial.asp

Bibliografía

Alonso-Arévalo, Julio. *ProCite 5.0: guía*. 2006. Consultado en: marzo, 2007.
http://eprints.rclis.org/archive/00007214/01/ProCite50.pdf

Armenteros-Vera, Ileana; **Alfonso-Sánchez, Ileana**. "Los gestores personales de bases de datos bibliográficas: conoce usted qué es y cómo se maneja el Procite". En: *Acimed*, marzo-abril, 2004, v. 12, n. 2. Consultado en: marzo, 2007.
http://scielo.sld.cu/scielo.php?pid=S1024-94352004000200006&script=sci_arttext

Bravo-Toledo, Rafael. "Gestores personales de bases de datos bibliográficas". En: *El profesional de la información*, 1996, octubre, n. 48. Consultado en: marzo, 2007.
http://www.elprofesionaldelainformacion.com/contenidos/1996/octubre/gestores_personales_de_bases_de_datos_bibliogrficas.html

Cascales-Salinas, Bernardo, et al. *El libro de Latex*. Madrid: Pearson Prentice Hall, 2006. ISBN 84-205-3779-9.

Codina, Lluís. "Reference Manager: herramientas para el trabajo intelectual". En: *El profesional de la información*, 2000, octubre, v. 9, n. 10, pp. 20-21.

EndNote: gestor de referencias bibliográficas, Windows versión X. 2007. Consultado en: abril, 2007.
http://www.ubu.es/biblioteca/servicios/endnote/MANUAL_EndNote.pdf

EndNote Web. Consultado en: abril, 2007.
http://www.myendnoteweb.com/EndNoteWeb/1.3/release/help/ENW/help.htm

EndNote X: user´s guide. Thomson ResearchSoft, 2006. Consultado en: abril, 2007.
http://www.endnote.com/support/helpdocs/EndNoteXWinManual.pdf

IDG. Consultado en: abril, 2007.
http://www.idg.es/rss.asp

ProCite: version 5. Berkeley: Institute for Scientific Information Research-Soft, 1999. Consultado en: abril, 2007.
http://www.procite.com/support/docs/ProCite%205%20Manual.pdf

Reference Manager v. 11: guía de usuario. 2006. Consultado en: abril, 2007.
http://www.biblioteka.ehu.es/p006-8858/es/contenidos/informacion/guias/es_ind/adjuntos/RM11_EHU.doc

Reference Manager 11. [S.l.]: Thomson ResearchSoft, 2004.

RefWorks: manual de usuario. 2006. Consultado en: abril, 2007.
http://www.ucm.es/BUCM/servicios/doc5270.pdf

***Emilio Duarte-García**, Universidad del País Vasco, Biblioteca Campus de Álava "Koldo Mitxelena", C/ Nieves Cano, 33–Apdo. 138, 01080 Vitoria-Gasteiz*
emilio.duarte@ehu.es

31th ELAG Seminar sobre la biblioteca 2.0

Por **Sílvia Redondo**

Redondo, Sílvia. "*31th ELAG Seminar* sobre la biblioteca 2.0". En: *El profesional de la información*, 2007, noviembre-diciembre, v. 16, n. 6, pp. 657-660.

DOI: 10.3145/epi.2007.nov.13

Sílvia Redondo Iniesta *es licenciada en historia por la Universitat de Barcelona y en documentación por la misma universidad, donde también cursa el doctorado y ha obtenido el Diploma de Estudios Avanzados. Es bibliotecaria del CRAI de la Universitat de Barcelona y profesora asociada de la Facultat de Biblioteconomia i Documentació de dicha universidad.*

Introducción

EL *31TH ELAG LIBRARY SYSTEMS SEMINAR* tuvo lugar en Barcelona entre los días 9 y 11 de mayo, y reunió a 120 participantes vinculados al mundo de las bibliotecas y las tecnologías de la información y la comunicación (TIC) de más de 20 países europeos.

Este evento anual fue organizado en esta ocasión por el *Consorci de Biblioteques de les Universitats Catalanes* (*CBUC*), la *Universitat de Barcelona* (*UB*), la *Universitat Autònoma de Barcelona* (*UAB*) y la *Universitat Pompeu Fabra* (*UPF*) y se desarrolló en el Edificio Histórico de la *UB* y en la recién inaugurada *Biblioteca de Filosofía, Geografía e Historia* de la misma universidad.

El *ELAG Seminar* ha tenido como lema central de esta edición la *Library 2.0* (biblioteca 2.0), aunque también se trataron otros aspectos como el o*pen access*, las publicaciones digitales o la interoperabilidad entre sistemas. El seminario se estructuró en tres bloques a lo largo de los tres días: los *papers,* los *workshops* y los *progress reports*, que se hicieron el último día.

Los *papers* son conferencias en donde pueden asistir todos los participantes. Con el fin de dar a conocer el ámbito bibliotecario de cada país en donde se realiza el seminario, se combinan intervenciones de carácter local con otras más generales relacionadas con el lema principal del evento. Por su parte, los *workshops* son talleres o grupos de trabajo coordinados por un especialista y destinados a fomentar la comunicación, el intercambio de ideas y la participación de los asistentes. Aquí, los participantes que se han inscrito dialogan sobre diferentes aspectos y comparten sus experiencias con el fin de llegar a unas conclusiones generales. Por último, los *progress reports* son informes que envían algunos participantes y en donde se hace una exposición de la actividad de la institución a la que pertenecen. Estos documentos se cuelgan en la página web que se crea cada año y se comentan durante el seminario.

Página web del 31th ELAG Seminar

"Los seminarios que organiza *ELAG* son una ocasión excepcional para reunirse con especialistas de diferentes nacionalidades e intercambiar experiencias y opiniones"

Asistentes al seminario. Foto Ramon Ros

Papers

De los once *papers* presentados, tres tuvieron carácter local, con temáticas diversas: los consorcios de bibliotecas, los repositorios y el uso de las publicaciones periódicas en formato digital. **Lluís Anglada**, director del *Consorci de Biblioteques Universitàries de Catalunya*, en su exposición *Twenty-five years of library automation in Catalonia*, pasó revista al proceso de automatización que han seguido las bibliotecas catalanas en los últimos 25 años y presentó el estado actual de este proceso, incidiendo en la importancia de la cooperación entre bibliotecas y la creación de un único catálogo.

Reme Melero, del *Instituto de Agroquímica y Tecnología de Alimentos* de Valencia, se refirió a los depósitos digitales en acceso abierto en España en *Open access institutional repositories: the case study of Spain*. En su estudio destacó el rápido crecimiento de estos, particularmente entre los años 2004 y 2005, y la utilización de los programas *Dspace* y *ePrints* para implantarlos. En España, los depósitos digitales más frecuentes son institucionales y suelen contener, entre otras tipologías documentales, contribuciones a congresos, tesis, trabajos de licenciatura, *pre-prints* y *post prints.*

El último día **Àngel Borrego**, profesor de la *Facultat de Biblioteconomia i Documentació de la Universitat de Barcelona*, presentó las conclusiones obtenidas a través de diversos estudios realizados sobre el uso de las revistas digitales en las universidades catalanas (*E-journal usage studies at Catalan academic libraries).*

De los *papers* de carácter general podemos destacar especialmente dos, el de **Maja Zumer**, de la *Universidad de Ljubljana,* y el de **Herbert Van de Sompel**, investigador de *Los Alamos National Laboratory Research Library*. **Zumer** presentó una conferencia que relacionaba *Amazon* y los catálogos de bibliotecas: *Amazon: competition or complement to OPACs*. Explicó como diversos estudios muestran las dificultades que tienen muchos usuarios para utilizar los catálogos de bibliotecas, debido especialmente a que las interficies de búsqueda son muy poco amigables. Esto provoca que algunos utilicen por ejemplo *Amazon* para buscar el ISBN y se sirvan de esta información para completar la búsqueda en el Opac. **Zumer** destacó las ventajas de la utilización del conocido portal de libros frente a los catálogos tradicionales: ofrece una interficie de búsqueda simple, muestra la cubierta del documento, permite hacer recomendaciones a los usuarios y ofrece la posibilidad, si existe, de enlazar con el texto completo. Si bien es cierto que algunas bibliotecas ya han empezado a aplicar algunas de estas funciones, la mayoría de centros siguen esperando para adaptar sus catálogos a las necesidades de sus usuarios. Tal y como insistió la ponente, sería necesario dirigir los cambios poco a poco en esa dirección y perder el miedo a equivocarse.

Van de Sompel centró su exposición en la descripción de los trabajos de *ORE*, el nuevo proyecto del equipo de *Open Archive Initiative* (*OAI Object Re-Use and Exchange: Moving interoperability from the metadata to the resource level). ORE* surgió a raíz del rápido aumento de repositorios, especialmente académicos, y de la gran cantidad de documentos que contienen. La idea de aprovechar estos repositorios y la documentación que hay en ellos hizo que el equipo de *OAI* se planteara la posibilidad de aprovechar esta situación para crear una marco de comunicación digital internacional de documentos. Este marco facilitaría el flujo de los objetos que contienen los repositorios

Van de Sompel durante su exposición. Foto Ramon Ros

y las diferentes partes de éstos (texto, imágenes, etc.). Así, se trata de crear un protocolo que permita no sólo intercambiar los metadatos de los documentos que contienen los repositorios sino intercambiar los mismos documentos y/o las diferentes partes de éstos.

El resto de *papers,* con autores de diferentes nacionalidades, trataron temas como: *Moving towards a service-oriented architecture* (**Erlend Gutteberg**), *Electronic publishing and institutional archives: utilising open-source software* (**Ellen Royneberg**), *Accessing library material via Google and other websites* (**Janifer Gatenby**), *RDA (Resource Description and Access)* (**Gordon Dunsire**), *A futuristic view of knowledge and information management* (**Sue Mcnight**) y *The DELOS reference model for digital libraries* (**Vittore Casarosa**).

Workshops y progress reports

Este año se ofreció a cada participante la posibilidad de participar en un *workshop*, en función de su especialidad y sus intereses. Los temas tratados en estos talleres fueron asimismo variados. *Google* y *Amazon* volvieron a ser protagonistas en el taller *Google, Amazon and libraries.* Aquí los participantes debatieron temas como la confianza que deben tener o no los bibliotecarios en *Google* y *Amazon,* la necesidad o no de las bibliotecas de exponer sus fondos en el máximo de interficies posibles o las nuevas formas de catalogación que podrían implantarse.

La preservación también tuvo cabida en el *Seminar* de este año en el taller *Preservation of digital content* donde se habló de los estándares utilizados (*MIX*, *PREMIS*, *OAIS*, *PDF/A*), en la calidad, el control de la digitalización y las normativas existentes.

Los *workshops* con más participantes fueron los relacionados con sistemas y servicios 2.0, como: *"Blogs, wikis,..."* y "*Library 2.0: what is in a name?"*. En el primero, coordinado por **Ramon Ros**, del *Consorci de Biblioteques Universitàries de Catalunya*, asistieron gran parte de los asistentes nacionales y se trataron temas relacionados con la aplicación en bibliotecas de algunas de las herramientas y tecnologías relacionadas con la web 2.0.

Por su parte, **Ole Husby** de *Bibsys,* el sistema de automatización de bibliotecas de Noruega, se encargó de coordinar el taller sobre el tema principal del *Seminar*, la *Library 2.0.* Los participantes en este grupo de trabajo profundizaron en ese concepto y plantearon algunas cuestiones que dejaron abiertas para posteriores reflexiones. Así, surgieron discusiones como la con-

Workshop Blogs, wikis,... Foto Ramon Ros

El grupo *ELAG*

ELAG (*European Library Automation Group*) es una organización formada por profesionales europeos interesados en las aplicaciones de las tecnologías de la información y la comunicación (TIC) en las bibliotecas. Organiza desde hace más de 30 años seminarios anuales, que tradicionalmente se han caracterizado por tratar temas relacionados con las tecnologías en centros bibliotecarios, aunque últimamente la temática se ha diversificado más y se han explorado nuevos ámbitos como la web semántica, el *open access*, la interoperabilidad de sistemas o las nuevas herramientas bibliotecarias.

fianza que deben tener los profesionales en los usuarios para permitirles participar o la definición de los agentes implicados. Para profundizar en los temas tratados el coordinador del grupo diseñó un blog[1] que utilizó como herramienta de trabajo y de comunicación en las diferentes sesiones y en la presentación general de los resultados que se hizo al final.

En el resto de *workshops* se trabajaron otros temas como la utilización de etiquetas o *tags* (*Social tagging/indexing (User participation*)), el acceso abierto (*Open-source software: pros and cons)* o el *E-learning* (*E-learning and its effects on libraries).*

Por su parte, los *progress reports* fueron presentados por **Ramon Ros** y, aunque hay que lamentar que no hubiera demasiada participación, éste se encargó de realizar algunas preguntas a los responsables de los informes que sirvieron para generar diálogo y discusión.

Conclusiones

Los seminarios que organiza anualmente el grupo *ELAG* son una ocasión excepcional para reunirse con especialistas de diferentes nacionalidades e intercambiar experiencias y opiniones. Las distintas ubicaciones geográficas del seminario permiten a los asistentes conocer, mediante los *papers* locales, contextos bibliotecarios y experiencias que, de otro modo, difícilmente hubieran conocido.

Después de escuchar las diferentes aportaciones de los *papers,* los *progress reports* y los *workshops*, una de las cuestiones que cobra más fuerza es que la biblioteca 2.0 ofrece a los profesionales y a las instituciones a las que pertenecen una gran oportunidad para mejorar sus servicios y su reconocimiento. Aunque actualmente ya hay diversas bibliotecas que empiezan a utilizar herramientas características de la web 2.0, como los blogs, en este *Seminar* se han señalado algunos aspectos cruciales que están frenando su avance como son la falta de confianza en el usuario y el miedo a la manipulación externa del catálogo.

Es necesario por tanto que los profesionales empecemos a adaptarnos a este nuevo contexto, confiando en los usuarios y en sus aportaciones y planteándonos la posibilidad de cambiar y de arriesgarnos con nuevas propuestas, ya que, como apostilló **Maja Zumer**: *"It is better to make an occasional wrong step than stand still and disappear into disuse"*.

"Es necesario que los profesionales empecemos a adaptarnos a este nuevo contexto, confiando en los usuarios y en sus aportaciones"

Para profundizar en los temas tratados en este seminario, se puede consultar desde mediados de octubre en los *Proceedings of the 31st Library Systems Seminar: ELAG 2007: Library 2.0*[2] toda la documentación generada referente a los *papers, workshops* y *progress reports.* Asimismo, también se pueden ver fotografías del acontecimiento y consultar otras informaciones de interés.

Por último, se puede anunciar ya que el próximo *ELAG Seminar*, el número 32, tendrá lugar entre los días 14 y 16 de abril de 2008 en el *University and Research Centre Library* de Wageningen, (Holanda), con el lema central "*Rethinking the Library"*. Actualmente ya se puede consultar el programa en su web[3], por lo que es un buen momento para empezar a preparar nuestras agendas y reservarnos esos días para la asistencia al próximo *Seminar* donde tendremos ocasión de replantearnos una nueva biblioteca.

Referencias

1. *http://ws5e.wordpress.com/*

2. *http://elag2007.upf.edu/*

3. *http://library.wur.nl/elag2008/*

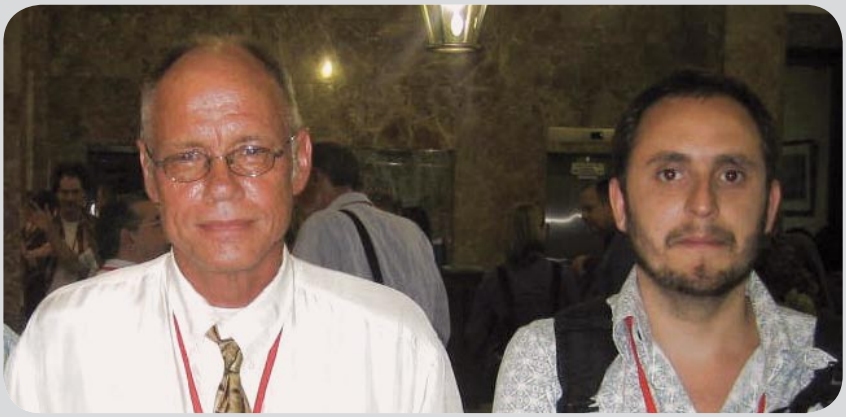
***Sílvia Redondo**, Universitat de Barcelona, Facultat de Biblioteconomia i Documentació*
redondosil@gmail.com

Diseño de un sistema de información y evaluación científica (Doctoral Thesis by Daniel Torres-Salinas)

Por **Henk F. Moed**

Moed, Henk F. "Diseño de un sistema de información y evaluación científica (Doctoral Thesis by Daniel Torres Salinas)". En: *El profesional de la información,* 2007, noviembre-diciembre, v. 16, n. 6, pp. 661-663.

DOI: 10.3145/epi.2007.nov.14



Henk F. Moed y Daniel Torres Salinas

QUANTITATIVE STUDY OF SCIENCE AND TECHNOLOGY is a rapidly developing field, and its development is closely linked to a number of general tendencies in scholarship throughout the world. National governments and research organisations and institutions need systematic evaluations to optimise their research allocations, re-orient their research support, justify research organisations, restructure research in particular fields, or augment research productivity.

Evaluative bibliometrics is a subfield of quantitative science and technology studies, intended to construct indicators of research performance from a quantitative analysis of scholarly documents. Citation analysis, one of its key methodologies, assesses the contributions scholars make in their research publications to the advancement of scholarly knowledge. During the past decades, numerous studies applied and further developed citation analysis in the assessment of scientific research performance. They reached the conclusion that bibliometric indicators can assist in building up insight into the quality of scholarly work under evaluation and in forming a judgment, and hence constitute a research evaluation *tool*, provided that they have a high level of sophistication and are derived from accurate bibliometric data.

This notion plays a key role in the dissertation of **Daniel Torres-Salinas**, entitled *Design of a system of scientific information and evaluation. A scientometric analysis of the research activities at the University of Navarra in the health sciences (1999-2005)*; the Spanish title is: *Diseño de un sistema de información y evaluación científica. Análisis ciencimétrico de la actividad investigadora de la Universidad de Navarra en el área de ciencias de la salud. (1999-2005*). Supervisors were **Evaristo Jiménez-Contreras** and **Emilio Delgado-López-Cózar** at the *University of Granada*. It presents the development of an information system on research activities in a scientific-scholarly institution that aims at providing a useful tool in research evaluation and management. The author carried out his work along the following five lines:

– The theoretical specification of such an information system and the type of science indicators included therein, in close interaction with potential users of the system.

– Its technical realization, including the development of various types of data entry processes, -partly manual, partly automated–.

– The actual application of the system to a particular case: the *University of Navarra*, a private Spanish university specializing in health sciences.

– An analysis of the results in terms of their implications for research performance at the *University of Navarra*.

– A discussion of his empirical findings from the perspective of more general research issues in the field of quantitative science studies and research assessment.

Life sciences departments at the *University of Navarra* constituted the object of the study. Data were collected from this university's internal databases, especially those on grant-funded projects and on scientific personnel, and from two major databases of scientific literature: the *Web of Science* published

by *Thomson Scientific* (formerly the *Institute for Scientific Information*) and *Scopus*, compiled by *Elsevier*.

The system includes both input and output indicators. The principal aspects of the output of scientific work covered by the system are: production, based on the number of published articles; impact or visibility at the international research front, derived from citation counts; and scientific collaboration, as reflected in co-authorship in scientific publications. Main aspects at the input side are the research capacity, measured by the number of active researchers, and the amount of external funding. Indicators were calculated at various levels of aggregation: for the institution as a whole and per department, thematic subfield and individual researcher.

The dissertation presents numerous interesting methodologies, results, observations and conclusions. The data tables and especially the figures are of very high quality. Below I highlight the study's outcomes and conclusions of a more general nature that relate to three issues currently receiving much attention in the field of quantitative science studies. The first deals with scientific literature databases: how do indicators derived from *Thomson Scientific's Web of Science* (*WoS*) compare to those obtained from *Elsevier's Scopus*? A second issue is: does scientific collaboration pay? More specifically: how does the citation impact of research articles resulting from collaboration –especially international collaboration– compare to that of non-collaborative papers? Finally, I address the analysis of the relationship between 'input' and 'output' of scientific research activity.

1. *Web of Science* versus *Scopus*

One of the objectives of Daniel Torres Salinas' study was to obtain insight into differences between citation impact indicators based on WoS data and those derived from Scopus. Scopus is a new multi-disciplinary citation index published by Elsevier, covering some 15.000 sources, mostly in science, technology and medicine. Although a number of recent studies compared Scopus and WoS, only a very few compared citation counts derived from the two databases. Torres Salinas' dissertation presents such a comparison. His comparative analysis of Scopus and WoS included citation counts for about 2.300 articles published from the University of Navarra and included in the WoS. Papers in Scopus journals not covered by the WoS were not taken into account. It was found that Scopus has almost 15 per cent more citations to these papers than the WoS.

Focusing on the largest subfields in terms of number of published articles, the comparative percentage of citations was somewhat higher in neurosciences, neurology and neurochirurgy, hematology and hematotherapy, preventive medicine and public health, and endocrinology and nutrition, and below the average (but nevertheless above zero) in internal medicine, gen therapy, oncology, biochemistry and molecular biology, genetics, and microbiology and parasitology. The ranking of departments based on *Scopus* citation counts was similar to that based on *WoS* citation data. Out of 50 departments, 20 were at the same position, 17 were higher in the *Scopus* ranking (13 by one position only), and 13 were lower (7 by one position).

Torres-Salinas makes the significant observation that, even though *Scopus* covers about 6.000 (or 65 per cent) more journals than the *WoS*, the citation counts to the papers published from the *University of Navarra* derived from *Scopus* are only some 15 per cent higher than those based on the *WoS*. He suggests that this discrepancy is due to differences in citation circuits between core and peripheral journals, assuming that *Scopus* contains more peripheral journals than the *WoS*.

As a rule, the use of multiple databases provides a more complete picture; more insight is needed into differences in coverage between *Scopus* and *WoS* and in citation patterns between core and peripheral journals. The outcomes of the study presented in **Torres Salinas**' dissertation constitute a proper starting point for further analysis, and point towards the possibility to carry out future bibliometric studies of citation impact based upon *Scopus* data only.

2. Does scientific collaboration pay?

The statistical relationship between scientific collaboration and citation impact, and the effects of collaboration on citation impact and vice versa, constitute important topics in scientometric research. In this research, collaboration is defined on the basis of (institutional) co-authorship. If all the authors of a paper are from one and the same institution, there is no formal institutional collaboration. If the authors are from two or more institutions located in the same country (and if no author from a foreign institution is involved), the collaboration is said to be national, while if the authors are from institutions located in two or more countries, the collaboration is defined as international.

Torres Salinas found for the *University of Navarra* that 26 per cent of its publication output resulted from international collaboration, 20 per cent from national collaboration, and 56 per cent did not result from collaboration formalized in institutional co-authorship. He found that internationally co-authored papers had on average a higher citation impact than papers resulting from national collabo-

ration, and that in turn nationally co-authored papers had a higher average impact than that of papers involving no collaboration. This outcome is in agreement with those obtained in similar studies for other universities.

Interestingly, however, **Torres-Salinas** further expanded this analysis by determining for each paper emerging from the *University of Navarra* the position of the Navarrese author(s) in the author list. Assuming that the first and the last author tend to be the most important contributors to a paper (i.e., the first author is often the junior researcher who carried out most of the work, whereas the last author is the project supervisor), he categorized the collaborative papers into two subclasses: those with a first or last author from the *University of Navarra* and those for which the author(s) from this university occupied an intermediate position (e.g., second or third author of a paper written by four authors). These two subclasses contained about 60 and 40 per cent of the collaborative papers published from the *University of Navarra*, respectively. Similar ratios were found both for internationally and for nationally co-authored papers.

He found that the average citation impact of an internationally co-authored paper in which the *University of Navarra* contributes the first or last author is lower than that of papers in which authors from this university occupied an intermediate position in the author list. This outcome is consistent with recent studies that are based on the notion that in a statistical analysis of the effects of scientific collaboration and its relationship to citation impact, one should examine 'who is collaborating with whom' and also take into account the type of contribution an author or institution made to a collective effort. The outcomes presented by **Torres Salinas** are a clear illustration of this.

3. Response surface analysis

Scientific activity can be described in terms of an input-output model as a system with easily defined borders that transforms inputs, such as funding, research capacity and equipment, into outputs, such as publications or patents. The relationships that link inputs with outputs are complex. Therefore, tools must be applied that are capable of more complex modelling. **Torres Salinas** applied such a technique, called Response Surface Methodology. It emerged in the 1950s in chemical engineering in an attempt to construct empirical models able to find useful statistical relationships between all the variables making up an industrial system. In recent years it is being applied successfully in biology, medicine, and economics.

Using this methodology for a set of 22 research departments, the author analyzed the statistical relationship between a department's amount of human resources and funding on the one hand and the number of published articles on the other. In addition, he studied the relationship between the degree of collaboration and prestige of journals in which a department published its papers on the one hand, and the average citation impact of its papers on the other.

One may ask whether social processes as complex as the production of scientific knowledge or the scientific reward system are ruled by the same type of causality as chemical-technological processes. But the outcomes of the methodology are certainly of interest, and the conclusions significant. One of his findings is that departments that have a strong capacity to actively collaborate are capable of taking better advantage of the results of the research and tend to generate higher citation impacts. A second conclusion states that systems that have fewer human resources with better funding tend to be more productive than those with more human resources but less funding.

Daniel Torres-Salinas' dissertation has made substantial contributions not only to the design and technical realization of scientific information systems and to insight into the research performance of the *University of Navarra*, but also to the exploration of new data analysis techniques that potentially have a wider applicability and to a deeper understanding of key issues in the field of science and technology studies.

***Henk F. Moed**, Centre for Science and Technology Studies (CWTS), Leiden University, P. O. Box 9555, 2300 RB Leiden, the Netherlands.*
moed@cwts.leidenuniv.nl

Presentación de *Medes* (Medicina en español)

Por **Tomàs Baiget**

Baiget, Tomàs. "Presentación de *Medes* (Medicina en español)". En: *El profesional de la información*, 2007, noviembre-diciembre, v. 16, n. 6, pp. 664-665.

DOI: 10.3145/epi.2007.nov.15

***Javier Ellena**, presidente de la Fundación Lilly; **Pedro García Barreno**, catedrático de Fisiopatología y Propedéutica Quirúrgicas de la UCM; y **José Antonio Gutiérrez Fuentes**, director de la Fundación Lilly.*

EL 25 DE JUNIO DE 2007 la *Fundación Lilly* presentó la base de datos *Medes*, que recoge las referencias bibliográficas con resumen de 55 revistas médicas españolas. En acto presentación participó el catedrático Pedro García Barreno, miembro de la *Real Academia Española* y de la *Real Academia de Ciencias*.

En el momento de escribir esta nota *Medes* contiene 22.000 registros, que pueden consultarse gratuitamente desde el URL:

http://www.fundacionlilly.com/medes/home.htm

Leemos en la web:

"La iniciativa *Medes* tiene como objetivo contribuir a promoción de la publicación en revistas biomédicas españolas, así como favorecer su difusión nacional e internacional mediante diversas actividades de formación y divulgación.

Desarrolla la base de datos bibliográfica *Medes-Medicina en español*, con el propósito de poner a disposición del profesional sanitario, fundamentalmente de la atención primaria, una herramienta de consulta centrada en aspectos de la práctica clínica, mediante una rigurosa selección de contenidos.

En el entorno de la iniciativa *Medes* se celebra una Jornada anual en la que se debaten temas relacionados con la publicación biomédica en España[1]. Las *Jornadas Medes* son muy interesantes, pues en ellas participan los mejores especialistas españoles, tanto responsables de política científica y editores de revistas, como evaluadores de artículos, autores, bibliotecarios, documentalistas..., con lo cual se establecen debates muy vivos. En la siguiente dirección pueden verse las transparencias y los vídeos de las ponencias, altamente recomendables:

http://www.fundacionlilly.com/medes/jornadas.htm

Medes cuenta con un Comité Técnico que valida la selección de los contenidos de la base de datos y participa en la elaboración de los programas de las jornadas y un Consejo Asesor compuesto por reconocidos profesionales del mundo de la medicina, la documentación, la publicación y la información científico técnica, con funciones consultivas y de representación".

¿Ciencia en español?

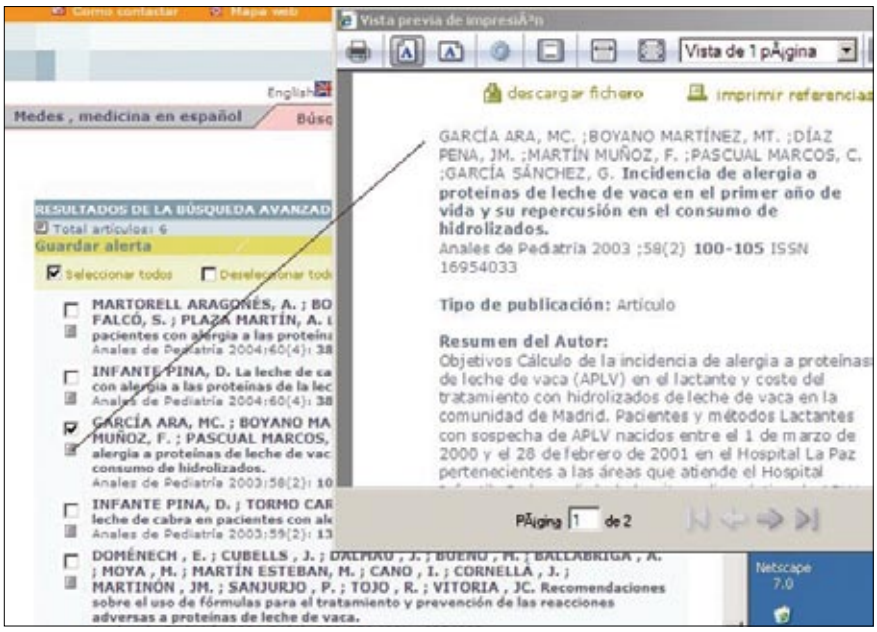
La respuesta a esta pregunta la encontramos en una publicación de la propia *Fundación*: "Parece indiscutible que el idioma de la ciencia es hoy el inglés, como lo fueron en otros tiempos el latín, el español o el alemán. Esta hegemonía favorece que los autores prefieran los libros y revistas en inglés para dar a conocer sus trabajos y que se ignoren las investigaciones que se realizan en otros países. No obstante cabría

Ángeles Flores, responsable del proyecto Medes

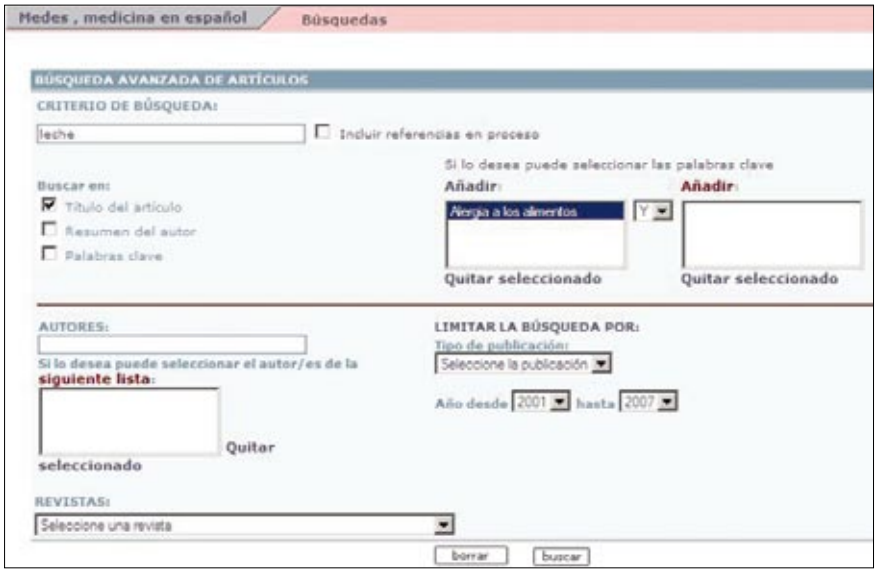
plantearse que ante trabajos que afectan directamente a la práctica de la medicina en nuestro entorno decisión de publicar sólo en revistas en inglés condena esas investigaciones al olvido: imposibles de entender y por ello carentes de interés para quienes no conocen el idioma".

1. **Roldán, Álvaro**. 2ª jornada de la Fundación Lilly sobre publicación médica en España (Hotel EuroForum, El Escorial, Madrid, 20 nov 2006). En: *El profesional de la información*, v. 16, n. 1, pp. 78-80.

http://eprints.rclis.org/archive/00009771/



Medes: Visualización de resultados



Medes: Búsqueda avanzada

2. **Baiget, Tomàs**. La publicación médica en España: crónica del curso (El Escorial, Madrid, 12 jul 2005). En: El profesional de la información, v. 14, n. 5, pp. 391-396.

http://eprints.rclis.org/archive/00006192/

Contacto: **Mª Ángeles Flores**; **David García-Cañadillas**
Fundación Lilly, María de Molina, 3 – 1º. 28006 Madrid.
Tel.: +34-917 815 070; fax: 917 815 079
fundacionlilly@lilly.com
http://www.fundacionlilly.com

***Tomàs Baiget**, baiget@sarenet.es*

Próximos temas centrales

Enero 2008	**Software libre para bibliotecas**
Marzo 2008	**Innovación en bibliotecas públicas**
Mayo 2008	**Presente y futuro de la profesión**
Julio 2008	**Libros electrónicos**
Septiembre 2008	**Información en la empresa**
Noviembre 2008	**Redes sociales**

Los interesados pueden remitir notas, artículos, propuestas, publicidad, comentarios, etc., sobre estos temas a:

epi@elprofesionaldelainformacion.com

online information 2007

WWW.ONLINE-INFORMATION.CO.UK/CONFERENCE

El profesional de la información

http://www.elprofesionaldelainformacion.com

Spanish Meeting Point

Stand 732, Conferencia Online Information
Londres, 4-6 diciembre 2007

Este año en la Online Information Conference de Londres, la más importante del mundo en materia de contenidos, se podrá visitar el stand Spanish Meeting Point, un punto de encuentro para los profesionales de la información hispanohablantes.

La financiación de este stand ha sido posible gracias a:

CIEPI
CENTRO INTERNACIONAL PARA LA INVESTIGACIÓN EN ESTRATEGIA Y PROSPECTIVA DE LA INFORMACIÓN

CINDOC

doc6
consultores en recursos de información

EBSCO
INFORMATION SERVICES

ProQuest
Start here.

Wolters Kluwer Health | Ovid

Agenda

4-6 de diciembre de 2007
INTL. ONLINE INFORMATION CONFERENCE
INFORMATION MANAGEMENT SOLUTIONS
Londres
Katherine Allen
katherine.allen@vnuexhibitions.co.uk
Lorna Candy
lorna.candy@vnuexhibitions.co.uk
http://www.online-information.co.uk/

7 de diciembre de 2007
FOURTH ANNUAL EUROPEAN EDITORIAL MANAGER USER GROUP MEETING
Oxford
Aries Systems Corporation
http://www.editorialmanager.com/homepage/emusergroup.html

11-15 de febrero de 2008
FIFTH INTL. SYMPOSIUM ON FOUNDATIONS OF INFORMATION AND KNOWLEDGE SYSTEMS (FoIKS)
Pisa, Italia
http://2008.foiks.org/

13-15 de febrero de 2008
VIII CONGRESO NACIONAL DE ANABAD. Memoria y tecnología.
Madrid
Mª del Pilar Gallego Cuadrado. *Federación Española de Asociaciones de Archiveros, Bibliotecarios, Arqueólogos, Museólogos y Documentalistas (Anabad)*
http://www.anabad.org

27-29 de febrero de 2008
CONF. INTL. BRECHA DIGITAL E INCLUSIÓN SOCIAL
Madrid
Instituto Agustín Millares de Documentación y Gestión de la Información, Univ. Carlos III de Madrid; Programa Sociedad de la Información y el Conocimiento (Prosic), Universidad de Costa Rica

30 de marzo-3 de abril de 2008
30TH EUROPEAN CONFERENCE ON INFORMATION RETRIEVAL
Glasgow, UK
BCS Information Retrieval
http://ecir2008.dcs.gla.ac.uk/

31 de marzo-1 de abril de 2008
CLSIG/EUSIDIC CONFERENCE
Londres
Commercial, Legal & Scientific Information Group, Cilip / Eusidic
oriole@legal-im.co.uk, johanvanhalm@cs.com
http://www.iclg.org.uk, http://www.eusidic.net

1-4 de abril de 2008
3RD INTL. CONF. ON OPEN REPOSITORIES
Southampton, UK
School of Electronics and Computer Science, University of Southampton
Leslie Carr, *lac@ecs.soton.ac.uk*
http://or08.ecs.soton.ac.uk/
http://openrepositories.org/2008/

18-20 de abril de 2008
AUKML CONF., INFORMATION PROFESSIONALS ADDING VALUE
Manchester, UK
Association of UK Media Librarians (AKUML)
http://www.aukml.org.uk/

21-23 de abril de 2008
FOURTH NORDIC CONFERENCE ON SCHOLARLY COMMUNICATION (NCSC 2008)
Lund, Suecia
Lund University Libraries
http://www.lub.lu.se/ncsc

22-24 de abril de 2008
STM Annual Spring Conference
Cambridge, Massachusetts
International Association of Scientific, Technical, and Medical Publishers
http://www.stm-assoc.org/stm-conference/

4-7 de mayo de 2008
4TH INTL. CONF. ON WEB INFORMATION SYSTEMS AND TECHNOLOGIES (WEBIST)
Funchal, Madeira, Portugal
Institute for Systems and Technologies of Information, Control and Communication / Universidade da Madeira
http://www.webist.org/

22-23 de mayo de 2008
11ES JORNADES CATALANES D'INFORMACIÓ I DOCUMENTACIÓ
Barcelona
Núria Castells. *Col·legi Oficial de Bibliotecaris-Documentalistes de Catalunya (Cobdc)*
Tel.: +34-933 197 675
jornades2008@cobdc.org
http://www.cobdc.org/jornades/11JCD/index.html

2-7 de junio de 2008
LIBRARIES IN THE DIGITAL AGE (LIDA) 2008
Dubrovnik, and Mljet, Croacia
Inter-University Centre
lida@ffos.hr
http://www.ffos.hr/lida/

11-13 de junio de 2008
E-DOCPA 2008. LA INFOSFERA EN LA INNOVACIÓN DE LA E-ADMINISTRACIÓN. EL CONOCIMIENTO COMO PROCESO MODERNIZADOR DE LA GESTIÓN DOCUMENTAL Y DE LAS RELACIONES CON LA CIUDADANÍA
Oviedo
Pilar Sánchez-Vicente. *Gobierno del Principado de Asturias. Consejería de Economía y Administración Pública.*
Tel.: +34-985 105 355; fax: 985 105 790

1-5 de julio de 2008
37TH LIBER
Estambul
Liga de Bibliotecas Europeas de Investigación
Ilkay Holt. *Koç University*
http://liber2008.ku.edu.tr

20-25 de julio de 2008
INTL. ASSOC. OF MUSIC LIBRARIES, ARCHIVES AND DOCUMENTATION CENTRES ANNUAL CONF.
Nápoles
http://www.iaml.info/en/node/399

10-15 de agosto de 2008
74TH IFLA GENERAL CONFERENCE
Québec
http://www.ifla.org/IV/ifla74/index.htm

18-21 de septiembre de 2008
3RD REFORMA NATIONAL CONFERENCE
El Paso, Texas
Reforma (National Association to Promote Library and Information Services to Latinos and the Spanish-Speaking)
Tel.: +1-360 264 23 69
Selina Gomez-Beloz, *selina-reforma@comcast.net*
http://www.reforma.org

Octubre de 2008
5º CONGRESO DE ARCHIVOS DE CASTILLA Y LEÓN
León
Asociación de Archiveros de Castilla y León (ACAL)
http://www.acal.es/Congresosjornadas/tabid/160/Default.aspx

15-19 de octubre de 2008
FRANKFURT BUCHMESSE
Frankfurt
Frankfurt Book Fair, Inc.
http://www.frankfurt-book-fair.com/

El profesional de la **información**

BOLETÍN DE SUSCRIPCIÓN

Deseo recibir todos los números de la revista EPI a partir del mes de enero del año

Suscripción: Institucional Personal

Nombre: Institución:

(Los suscriptores individuales no han de escribir ningún nombre de institución, sólo indicar la dirección particular)

Departamento: NIF institucional:

Dirección:

Código postal: Ciudad: País:

Teléfono: Fax: Correo-e:

Método de pago:

Tarjeta de crédito: VISA Master Card American Express

Titular de la tarjeta:

Número de tarjeta:

Caducidad (mm/aaaa):

Cheque nominativo en euros a nombre de El profesional de la información

Transferencia bancaria a la cuenta de La Caixa 2100 0818 93 0200745544

Enviar, fotocopiado o escaneado, el resguardo de la transferencia.
Las transferencias desde fuera de España deben hacerse a:
IBAN ES95 2100 0818 9302 0074 5544
BIC/Código Swift CAIXESBBXXX

Giro postal al apartado de correos 32.280 de Barcelona

Enviar, fotocopiado o escaneado, el resguardo del giro.

Domiciliación en cuenta bancaria

Entidad: Oficina: DC: Núm:

Titular de la cuenta:

Enviar el boletín relleno, por correo postal o electrónico, a:

Apartado 32.280
08080 Barcelona
España

suscripciones@elprofesionaldelainformacion.com

Teléfono de atención al suscriptor:

+34 609 352 954

Los precios para el año 2008 son los siguientes:

Suscripción anual Institucional: 147,2 € + 4% IVA = 153 €

Suscripción sólo online: 85 € + 4% IVA = 88,4 €

Individual: 75 € + 4% IVA = 78,00 €

Número suelto: 25 € + 4% IVA = 26 €
(gastos de envío fuera de España: 7,00 €)

Coste adicional de correo aéreo:
España: 00,00 €
Europa (menos España): 30,00 €
Américas y resto del mundo: 45,00 €

La suscripción a la revista se realiza por años naturales completos, es decir, desde el mes de enero del año que usted desee que comience su nueva suscripción

Boletín para comenzar o renovar online la suscripción: http://www.elprofesionaldelainformacion.com/suscripciones.html

Información para los autores

Todos los profesionales que lo deseen pueden remitir a la redacción de la revista **El profesional de la información** sus colaboraciones en forma de:

- Notas breves
- Trabajos más amplios sobre temas de fondo para la sección "Artículos".

El texto ha de enviarse en formato electrónico. Aparte, los materiales gráficos, en papel o ficheros gif, jpeg o tiff con unos anchos de entre 12 y 5,7 cm. y una resolución de 300 ppp.

El tamaño ideal de un estudio para la secció "Artículs" es de 4.000 palabras. En casos excepcionales pueden publicarse artículos de mayor extensión. Los trabajos de esta sección son aprobados según el sistema tradicional "peer review": al menos dos expertos en el tema, del Consejo Asesor de la revista y/o externos, deben dar el visto bueno antes de su publicación.

Los textos deben enviarse sin formatos especiales (títulos, secciones, subsecciones, pies de página, sangrías, tabulaciones, colores, etc.).

Los trabajos para la sección "Artículos" deben incluir: a) título en castellano, b) resumen en castellano de 100-150 palabras, c) 5-10 palabras clave en castellano, d) título en inglés, e) resumen en inglés de 100-150 palabras, f) 5-10 palabras clave en inglés, g) texto completo en castellano y h) nombre de los autores, lugar de trabajo y dirección de correo electrónico. **Han de ser inéditos.**

Se valorará especialmente que los trabajos sean concisos y precisos. Se ruega a los autores que eviten una excesiva retórica.

Las citas bibliográficas en el texto se realizarán de la forma: (Apellido, año). Las referencias bibliográficas, que se limitarán a las obras citadas en el texto, han de prepararse de acuerdo con el siguiente esquema:

Artículos de una publicación periódica:
Apellido, Nombre; Apellido2, Nombre2. "Título del artículo". En: Título de la publicación periódica, año, mes, v. [volumen], n. [número del ejemplar], pp. [págs. comienzo-final].

Ponencia presentada en un congreso:
Apellido, Nombre; Apellido2, Nombre2. "Título de ponencia". En: nombre del congreso, año, pp. [págs. comienzo-final].

Capítulo de una monografía:
Apellido, Nombre; Apellido2, Nombre2. Título del capítulo. En: Apellido, Nombre; Apellido2, Nombre2. Título de la monografía. Lugar de publicación: editor, fecha. ISBN [número].

Monografías:
Apellido, Nombre; Apellido2, Nombre2. Título del trabajo. Lugar de publicación: Editor, fecha. ISBN [número]

Recurso en línea:
Apellido, Nombre; Apellido2, Nombre2. Título del recurso. Consultado en: día-mes-año.

Dirección:

Las contribuciones se pueden enviar a la redacción de la revista o a cualquiera de los miembros del consejo de redacción.

El hecho de que un trabajo sea publicado en EPI no implica que la redacción se adhiera a las opiniones expresadas en él.

Redacción EPI:
Apartado 32.280
08080 Barcelona.
epi@elprofesionaldelainformacion.com

La redacción se reserva el derecho de adaptar los textos al estilo gramatical y literario de la revista.