ISKO. CAPÍTULO ESPAÑOL. CONGRESO (7º: 2005: BARCELONA) 33-50

L'accés multilingüe per matèries a articles de revista

Ernest Abadal, abadal@ub.edu
Assumpció Estivill, estivill@ub.edu
Jorge Franganillo, franganillo@ub.edu
Jesús Gascón, gascon@ub.edu
Josep Manuel Rodríguez Gairín, rodriguez.gairin@ub.edu

Universitat de Barcelona (Barcelona, Espanya). Departament de Biblioteconomia i Documentació

Resum

Es presenta una anàlisi dels diversos models d'utilització de tesaurus en serveis web per facilitar l'accés temàtic multilingüe a col·leccions digitals. A continuació, es descriu el cas de l'adaptació del *Tesauro de biblioteconomía y documentación* (elaborat pel CINDOC) per ser utilitzat en un portal d'articles de revista i que permet, a la vegada, una consulta multilingüe. Finalment, es conclou valorant l'eficàcia dels tesaurus per a la recuperació del contingut de col·leccions digitals de manera multilingüe.

Paraules clau

Temària, accés temàtic multilingüe, tesaurus multilingües, portals temàtics, traducció automàtica de consultes, equivalència lingüística, interoperabilitat semàntica, metadades, bases de dades d'articles de revista, Bireme, *DeCS*, *Labordoc*, *MACS*, Organización Internacional del Trabajo

1 Introducció

Els catàlegs de biblioteca i les bases de dades inclouen, normalment, recursos documentals escrits en una gran diversitat de llengües. Tanmateix, només acostumen a permetre la consulta en un sol idioma, que és el de l'àmbit lingüístic on es troba la base de dades.

Internet ha facilitat que els usuaris d'aquests catàlegs i bases de dades no siguin només de l'entorn geogràfic més proper, sinó que provinguin d'arreu del món i que, per tant, utilitzin per a la consulta llengües diverses. Així doncs, a part de disposar de la interfície de navegació —els menús, les opcions, els missatges, etc.— en el seu idioma, aquests usuaris agraeixen també el fet de poder consultar (accedir temàticament a) aquestes col·leccions de documents en la seva llengua.

La consulta s'ha de poder dur a terme a partir dels camps temàtics (matèries, resum, etc.) que constitueixen l'eix fonamental de la recuperació de la informació, i un usuari català hauria de poder utilitzar els termes de ma-

tèria o les paraules del resum en la llengua d'Ausiàs Marc; un usuari castellà, en la de Cervantes, i un d'anglès, en la de Shakespeare. Els resultats trobats (ja siguin referències o documents complets), però, estaran en l'idioma en què han estat escrits i només es disposarà dels camps temàtics (els descriptors i, a vegades, també els resums) en la llengua de l'usuari.

Aquesta qüestió ha estat tractada i desenvolupada especialment, en països plurilingües o bilingües (Suïssa, el Canadà, Bèlgica, Israel, etc.), per organitzacions supranacionals (com ara la Unió Europea i les Nacions Unides) i per organitzacions internacionals o multinacionals (com ara l'Organització per a l'Alimentació i l'Agricultura (FAO) o l'Organització Internacional del Treball). Els sistemes d'informació d'aquests països i d'aquestes organitzacions han de disposar de funcionalitats multilingües per a l'accés a la informació. De tota manera, aquesta tendència es va estenent cada cop més, ja que l'usuari agraeix especialment poder emprar la seva llengua materna en tots els àmbits i, per tant, també quan duu a terme consultes d'informació.

Així doncs, hi ha una necessitat real de promoure i desenvolupar sistemes que facilitin la realització de consultes d'informació a bases de dades en la llengua de l'usuari. Això contribueix a aproximar l'usuari als termes del llenguatge documental. La resolució d'aquest problema pot dur-se a terme, fonamentalment, de dues maneres: per mitjà de la traducció automàtica de les consultes (*cross-language text retrieval*), o per mitjà de la utilització de llenguatges documentals (*multilingual subject access*).

La primera d'aquestes vies (*cross-language text retrieval* o *multilingual text retrieval*) s'investiga sobretot des de l'àmbit de la informàtica, i es basa en el desenvolupament de sistemes automàtics de traducció de les consultes o d'expansió semàntica cap a termes en altres idiomes. Oard i Dorr (1996) i Oard (1997) han elaborat un ampli i detallat estat de la qüestió de tots els estudis i les investigacions dutes a terme sota aquesta orientació.

Els desenvolupaments que es presenten al CLEF (Cross-Language Evaluation Forum) o el treball d'Ogden i Davies (2000) en són un bon exemple. El CLEF és una reunió científica[1] que té per objectiu promoure la recuperació d'informació multilingüe i que es va crear per estudiar i avaluar tecnologies que permetin accedir a la informació en múltiples idiomes. La base de les recerques que es presenten és la traducció automàtica de les preguntes formulades al sistema per part dels usuaris.

Per la seva banda, Ogden i Davies (2000) han desenvolupat un sistema de recuperació de text en diversos idiomes (*cross-language text retrieval*) que permet recuperar documents en una llengua diferent a la de la consulta. La base és la traducció automàtica dels termes de consulta a altres idiomes que pot l'usuari dominar o llegir, tot partint de diccionaris especialitzats, i no pas generals.

La segona de les orientacions, l'accés temàtic multilingüe (*multilingual subject access*), inclou els sistemes que permeten consultar temàticament

1. Està coordinada per DELOS <http://www.delos.info>, una xarxa d'excel·lència en biblioteques digitals patrocinada pel 6è Programa marc de la UE, que té per objectiu el desenvolupament tecnològic de la nova generació de biblioteques digitals.

un fons determinat a partir d'un llenguatge documental multilingüe. És un model bastant diferent a l'anterior, ja que no hi ha traducció automàtica dels termes de consulta, sinó que es tracta de veure la concordança entre els termes introduïts per l'usuari i els que formen part del llenguatge documental multilingüe de què disposa el catàleg o base de dades.

El nostre text se centra especialment en aquesta darrera orientació, de la qual es presentaran els models bàsics de desenvolupament, de manera teòrica, però també analitzant alguns casos reals concrets, amb l'objectiu últim de valorar l'eficàcia dels tesaurus per a la recuperació del contingut de col·leccions digitals de manera multilingüe. En primer lloc, es mostrarà un breu estat de la qüestió de la utilització de llenguatges documentals per facilitar l'accés temàtic multilingüe a col·leccions digitals i se n'analitzaran algunes de les experiències i dels desenvolupaments més destacats. En segon lloc, es presentarà el desenvolupament realitzat al portal *Temària* (www.temaria.net) per facilitar la consulta multilingüe d'articles de revista de biblioteconomia i documentació. El projecte ha partit de l'adaptació del *Tesauro de biblioteconomía y documentación* (MOCHÓN 2002).[2]

2 ESTAT DE LA QÜESTIÓ

L'accés temàtic multilingüe a una col·lecció de documents es pot aconseguir de dues maneres diferents: partint de tesaurus multilingües o desenvolupant sistemes d'equivalències (mapatge). En el primer cas, es disposa d'un llenguatge documental comú i, en el segon, el que es fa és establir equivalències entre termes de diferents llenguatges documentals. Aquesta darrera opció no està, ni de bon tros, tan implementada com l'anterior, ateses les dificultats que comporta.

El resultat d'utilitzar qualsevol d'aquests dos models és dotar la base de dades de, com a mínim, tres aspectes o funcions multilingües: la cerca (o sigui, poder introduir els termes de cerca en més d'un idioma), la visualització (això és, la possibilitat de veure els camps temàtics dels registres en diversos idiomes, i també poder escollir més d'una llengua per navegar pel tesaurus) i l'exportació de resultats (és a dir, disposar de prestacions per obtenir el resultat de la consulta en més d'un idioma).

A continuació es descriuen cada un dels models i se n'analitzen exemples concrets per veure com es duu a terme l'adaptació de tesaurus al web i la consulta multilingüe. Es faran comentaris sobre cada una de les tres funcions abans esmentades: la cerca, la visualització i l'exportació.

2.1 Utilització de tesaurus multilingües

L'aplicació de tesaurus multilingües a la consulta web és una opció adoptada per la majoria de catàlegs i bases de dades que ofereixen aquest ti-

2. Aquest llenguatge documental ha estat creat i desenvolupat pel CINDOC, i s'utilitza per a la indexació de la producció científica espanyola en biblioteconomia i documentació que està inclosa en la base de dades *ISOC*.

pus de prestacions. En aquest cas, l'usuari pot consultar la base de dades a partir d'un sol llenguatge documental que es troba en diversos idiomes. En molts casos, s'acostuma a partir d'una versió original en un idioma.[3] Dos exemples representatius d'aquest model són la versió multilingüe del *Medical subject headings* (MeSH), utilitzat pel MEDLINE i adaptat per la BIREME, i el tesaurus de l'Organització Internacional del Treball.

2.2 La Biblioteca Virtual en Salud i el DeCS, la BIREME[4]

El *DeCS* (*Descriptores en ciencias de la salud*, http://decs.bvs.br/E/homepagee.htm) és un tesaurus trilingüe —anglès, espanyol i portuguès— creat el 1986 per la BIREME a partir de la traducció i l'adaptació del MeSH (www.nlm.nih.gov/mesh/meshhome.html).[5] S'empra com a llenguatge únic i uniforme en el desplegament de processos d'indexació i de recuperació d'informació cientificotècnica en el marc de la LILACS, la xarxa Literatura Latinoamericana y del Caribe en Ciencias de la Salud. En aquest sentit, Rodríguez Camiño (1998) fa referència a l'aplicació del DeCS al Sistema Nacional de Información de Ciencias Médicas de Cuba (SNICM) a partir de l'any 1991 i a les accions necessàries per permetre que la xarxa nacional conegui, apliqui i difongui els canvis aplicats al vocabulari.

De fet, el DeCS és l'eina utilitzada en la indexació i recuperació d'informació en les diferents bases de dades especialitzades en medicina que formen l'anomenada Biblioteca Virtual en Salud (www.bireme.br/bvs/E/ebd.htm), com ara la LILACS i el MEDLINE.[6] Els conceptes s'organitzen en una estructura jeràrquica que acull més de 26.000 descriptors i que experimenta un creixement continuat (actualització anual). A banda dels termes originals del MeSH, han estat desenvolupades dues àrees específiques: salut pública i homeopatia. Per mantenir el mateix codi jeràrquic en els tres idiomes, la llista va ser preparada a partir dels descriptors en anglès. Així doncs, en espanyol i portuguès els descriptors no apareixen en ordre alfabètic dins de l'estructura jeràrquica del DeCS.

La consulta del tesaurus en línia permet accedir als termes per mitjà de l'índex alfabètic, el permutat o el jeràrquic, tot mantenint la possibili-

3. Això és el que es va fer, per exemple, amb *AGROVOC*, el tesaurus d'agricultura desenvolupat per la FAO, que va ser traduït al castellà, el portuguès, el xinès, l'àrab i el txec a partir de la versió original en anglès.
4. La Biblioteca Regional de Medicina (www.bireme.br) forma part de l'Organización Panamericana de la Salud (OPS). Va ser creada al Brasil l'any 1967, i té com a objectiu prioritari contribuir al desenvolupament de la salut per mitjà de l'enfortiment i l'ampliació dels fluxos d'informació en les ciències de la salut. Una de les seves creacions en aquest sentit és la Biblioteca Virtual en Salud (BVS, www.bireme.br/bvs/E/ehome.htm), entesa com a base del coneixement cientificotècnic en salut als països de l'Amèrica Llatina i del Carib. La BIREME també coordina el desenvolupament i l'actualització de la terminologia relacionada amb les ciències de la salut, i la difon en anglès, espanyol i portuguès, per mitjà del vocabulari *Descriptores en ciencias de la salud (DeCS).*
5. El MeSH va ser elaborat el 1963 per la National Library of Medicine.
6. A banda de la LILACS i el MEDLINE, el portal també incorpora deu bases de dades d'accés gratuït especialitzades en ciències de la salut.

tat de fer una cerca per paraula. D'aquesta manera accedim al registre del terme seleccionat, que inclou, entre altres informacions, la designació dels descriptors en els tres idiomes, sinònims en l'idioma emprat en la cerca i la o les categories en què s'inscriu el terme. Fátima Pellizzon (2004) explica amb detall les possibilitats de consulta que integra el tesaurus en línia.

D'altra banda, la consulta a la Biblioteca Virtual en Salud possibilita la configuració de l'idioma de la interfície (portuguès, anglès i castellà) i de la presentació dels resultats en qualsevol punt del procés de cerca (és una possibilitat de què també disposem en consultar el tesaurus en línia). El sistema també permet la consulta dels termes del tesaurus des de la base de dades. D'aquesta manera, l'usuari pot fer una cerca per matèries en qualsevol dels tres idiomes i visualitzar-ne els resultats en l'idioma de la interfície, agrupats per base de dades. Així, si es fa la prova d'una consulta per a un mateix concepte en qualsevol dels tres idiomes (per exemple, si es busca el descriptor *job satisfaction* / *satisfacción en el trabajo* / *satisfaçao no emprego*), es comprova que el resultat és el mateix en tots els casos. Cal destacar que, en accedir als resultats de la consulta, podem obtenir informació d'altres documents relacionats (en el sentit que contenen tots o algun descriptor coincident). Actualment (6 de març de 2005), aquesta darrera opció es troba, però, en procés d'avaluació.

2.3 La *Labordoc*, Organització Internacional del Treball

En el cas de la base de dades *Labordoc* (www.ilo.org/labordoc), s'ofereix accés multilingüe —francès, anglès i espanyol— a més de 350.000 referències bibliogràfiques de documents que tracten una gran varietat d'aspectes relacionats amb el treball (formes de vida sostenible, desenvolupament econòmic i social, etc.). Els documents han estat indexats per mitjà de l'*ILO thesaurus of labour, employment, and training terminology* (ILO 1998) des de l'any 1965.

Tal com apunta Davies (2003), de 1980 a 2002, el catàleg de la biblioteca de l'Organització Internacional del Treball (OIT) es gestionava amb el sistema *MINISIS*.[7] Aquest programa va possibilitar l'accés temàtic multilingüe a partir de l'*ILO thesaurus*. El funcionament d'aquest sistema es basava en la traducció automàtica dels termes (*online processing*). Així, els descriptors de matèries eren assignats en una única llengua —l'anglès—, i el procés de traducció de termes tenia lloc en el moment en què l'usuari feia la cerca en el catàleg. L'any 1997, el *MINISIS* va ser reemplaçat pel *Voyager*, un programa de disseny obert que va permetre posar en marxa un nou model per a l'accés multilingüe per matèries —Davies (2003) l'anomena *pre-processing model*.[8] En aquest cas, els descriptors són traduïts abans que l'usuari faci la consulta. De fet, la catalogació per matèries es

7. El *MINISIS* va ser creat per l'International Development Research Centre, una agència canadenca.
8. El *Voyager* és el programa que empren actualment moltes biblioteques, com la mateixa Library of Congress, per a la gestió dels catàlegs.

fa en anglès, i el programa estableix les equivalències i les incorpora de manera automàtica a l'etiqueta 650 del registre MARC.

En iniciar la consulta en la *Labordoc*, el sistema permet seleccionar l'idioma de la interfície —anglès, francès o espanyol. Tanmateix, aquesta operació no és possible durant el desenvolupament de les cerques.

Quan s'opta per la cerca simple i es busca una matèria en un dels tres idiomes possibles, com a resposta s'obté una llista de descriptors ordenats alfabèticament, amb la indicació del nombre de documents relacionats i del tesaurus d'on s'extreu cada descriptor en funció de la versió lingüística del terme: *Tesauro OIT* (per a l'espanyol), *Thesaurus BIT* (per als termes en francès) i *ILO thesaurus* (per als termes en anglès).

És destacable la informació que hi incorpora el registre MARC i que fa possible l'accés multilingüe (l'etiqueta 650 i una etiqueta local 9XX, amb els descriptors en les tres llengües).

Extracte del format marc del terme *labour contract*:

650 17 |a labour contract |2 ilot
650 17 |a contrat de travail |2 tbit
650 17 |a contrato de trabajo |2 toit

905 1_ |a labour contract
906 1_ |a contrat de travail
907 1_ |a contrato de trabajo

La llista de descriptors que s'obté en fer una cerca per matèria és l'índex alfabètic de matèries de la base de dades, que s'inicia amb el terme que ha estat objecte de cerca i que continua indefinidament, incorporant termes en les tres llengües emprades en la indexació.

Cal destacar que la *Labordoc* no inclou un apartat que permeti accedir al tesaurus. La cerca de termes del tesaurus només és possible des del camp *Subject* de la cerca simple de la base de dades.

Pel que fa a la presentació dels resultats, quan es fa una cerca per autor o per matèria, el resultat apareix en forma de llista de descriptors. De manera habitual, en gairebé tots els registres, hi apareix la icona «Más información / More info / Plus d'informations». Aquesta opció proporciona accés a suggeriments de noms i temes relacionats, notes d'aplicació, referències al terme acceptat, etc., i és una eina de suport valuosa per a la recuperació de la informació.

2.4 Equivalències (interoperabilitat)

L'establiment d'equivalències, interoperabilitat o mapatge[9] fa referència a la possibilitat de consultar de manera simultània diversos fons que han estat indexats amb llenguatges documentals diferents (l'idioma pot ser una d'aquestes diferències, però no és pas l'única). En aquest cas, l'èmfasi es posa a desenvolupar sistemes d'equivalències (mapatge) entre els

9. Clavel-Merrin (2004) ho anomena també *linking*.

termes de diferents llenguatges documentals ja existents. Martin Doerr (2001) tracta abastament d'aquesta qüestió i presenta la definició següent del terme:

> We regard thesaurus mapping as the process of identifying terms, concepts and hierarchical relationships that are approximately equivalent. It is a central process for merging thesauri, metathesaurus and cross-concordance construction, and thesaurus switching.

Doerr també assenyala quins són els problemes principals que cal afrontar en el procés de facilitar la interoperabilitat entre tesaurus. Són els següents: diferent ús dels termes, diferència en l'abast (un dels tesaurus pot ser més específic que els altres), diferències semàntiques i diferències en l'establiment de relacions semàntiques (les dependències jeràrquiques dels termes). Un dels tipus de mapatge que analitza és el realitzat amb tesaurus multilingües, tot i que aquest no és l'únic model descrit.

Chan i Zeng (2002) van publicar un article de caràcter global que feia un repàs general de les experiències d'interoperabilitat entre vocabularis temàtics i llenguatges documentals. Les possibilitats d'interoperar es refereixen a l'ús de diversos idiomes i també a la utilització de diversos llenguatges de classificació i d'indexació. Les autores assenyalen exemples en què s'apliquen o es desenvolupen sistemes d'interoperabilitat, amb alguns casos que fan referència a llenguatges documentals multilingües. En un treball posterior més extens (Zeng 2004), però de característiques similars, revisen aquesta mateixa qüestió.

Un dels exemples d'aquest model és el projecte *MACS* (*Multilingual access to subjects*) (Clavel-Merrin 2004; Landry 2000), que s'analitza a continuació.

2.5 El *MACS* (Multilingual access to subjects)

El *MACS* va ser creat el 1997 com a resposta a un encàrrec de la Conference of European National Libraries (CENL) amb l'objectiu de trobar solucions a l'accés multilingüe per matèries a bases de dades bibliogràfiques. Amb aquesta intenció es va crear un grup de treball sota el lideratge de Cobra+ (Computerised Bibliographic Record Actions), per tractar el tema en relació amb les biblioteques nacionals. Entre les aproximacions possibles, es va optar per la idea d'establir relacions entre diferents llenguatges d'encapçalaments de matèries.

El projecte *MACS* pretén proporcionar accés multilingüe per matèries —en anglès, francès i alemany— a diferents catàlegs simultàniament: el catàleg de les biblioteques nacionals de Suïssa (Swiss National Library), França (Bibliothèque nationale de France), Regne Unit (British Library) i Alemanya (Deutsche Bibliothek). El *MACS* parteix del convenciment que és possible la cerca multilingüe gràcies a la creació d'enllaços d'equivalència entre els tres llenguatges d'indexació (llistes d'encapçalaments) emprats a les biblioteques implicades en el projecte: *SWD* (Alemanya),

RAMEAU (França) i *LCSH* (Regne Unit). Tal com apunta Landry (2000), es va considerar que els llenguatges segueixen construccions i principis d'aplicació similars, i que la millor opció era partir dels llenguatges d'indexació ja existents, fonamentalment per tres motius: les biblioteques ja havien emprat temps i esforços considerables en la creació i el manteniment dels llenguatges documentals; els llenguatges d'indexació creats actualment ja proporcionen *de facto* accés per matèries a milions de documents, i la traducció seria molt costosa i requeriria que algunes de les llistes d'encapçalaments fossin abandonades.

Per comprovar-ne el funcionament, l'empresa danesa Index Data i la biblioteca neerlandesa Tilburg Universiteit Bibliotheek van crear un prototip (http://laborix.uvt.nl/prj/macs/prototypee.html) que conté una porció de les dades dels llenguatges d'indexació i de les bases de dades de les biblioteques nacionals participants en el projecte. El prototip (creat entre el 2001 i el 2002) ha seleccionat els encapçalaments dels tres llenguatges d'indexació en els àmbits del teatre i de l'esport, així com els encapçalaments més emprats per indexar documents a la biblioteca nacional francesa (500 termes extrets de la base de dades *RAMEAU* i els termes equivalents en els altres llenguatges d'indexació); un total de 3.000 encapçalaments i 30.000 registres bibliogràfics de les quatre biblioteques que participen en el projecte. Les dades presents al prototipus estan codificades en els formats MARC21, MAB/PICA i UNIMARC.

El *MACS* integra dues interfícies: l'anomenada *search interface*, que permet als usuaris buscar per encapçalament i recuperar els registres bibliogràfics presents a les tres biblioteques nacionals (per fer-ho possible s'empra el protocol Z39.50), i la *link management interface*, disponible des de 2001 i amb accés obert només com a lectura. Aquesta darrera interfície es va crear amb la intenció de permetre l'accés a les biblioteques que participen en el projecte i fer possible el manteniment de les relacions entre els termes dels diferents llenguatges. La interfície està protegida, i l'accés i la modificació dels encapçalaments emprats en la indexació als diferents llenguatges només és possible per mitjà d'un identificador i una contrasenya (a banda, existeixen diferents nivells d'accés). El sistema es basa en el principi de l'anomenada *federative management*, de manera que cada llenguatge d'indexació és autònom, i cada entitat col·laboradora és responsable de la gestió de les relacions del seu propi llenguatge —no hi ha un editor central. Els *partners* poden fer propostes sobre les parts implicades en una relació, però aquestes han de ser confirmades en darrer terme pel *partner* encarregat del manteniment del llenguatge afectat. Landry (2003) proporciona algunes pautes sobre els permisos d'accés a la modificació de termes i relacions per part dels agents implicats.

En la *link management interface*, la pantalla amb els resultats de la cerca mostra les relacions entre els encapçalaments dels diferents llenguatges documentals a partir del terme de cerca. En alguns casos s'estableix més d'una relació. Així, per exemple, en el cas de «jumping» (*LCSH*), s'estableixen dues relacions: una amb «saut en hauteur» (*RAMEAU*) i «hochsprung» (SWD), i una altra amb «sauts (athlétisme)» (*RAMEAU*) i «sprung» (SWD). Emprant l'opció «View link», obtindrem informació so-

bre quan van ser creades o modificades les relacions entre termes i qui va fer les modificacions.

L'usuari té la possibilitat de fer la consulta a una biblioteca nacional en particular i emprar una llista d'encapçalaments diferent a la que empra la biblioteca a la qual formula la consulta. També es pot seleccionar una de les llistes d'encapçalament de matèries (*RAMEAU, LCSH, SWD*) i interrogar un o més catàlegs amb el terme escollit. Cal dir que la possibilitat de seleccionar diferents termes alhora i llançar la consulta augmenta la capacitat de recuperació. Aquest és, de fet, un dels punts forts del recurs.

D'altra banda, els formats de presentació dels resultats a la base de dades bibliogràfica no són gaire amigables i ofereixen poques dades sobre els documents referenciats. També es troba a faltar una estructura de cerca jeràrquica que permeti el rastreig de termes i una visió global de les possibilitats d'accés per matèries del recurs.

Cal dir que el prototip encara no proporciona accés directe als catàlegs de les biblioteques participants. Tampoc no conté tots els conceptes dels llenguatges d'indexació, i de moment no inclou referències entre termes.

3 El cas de *Temària*

Temària és un portal d'articles de revistes espanyoles de Biblioteconomia i Documentació. De moment l'aplicació conté les metadades Dublin Core (DC) dels articles publicats a les revistes *Anales de documentación*, *BiD: tex-*

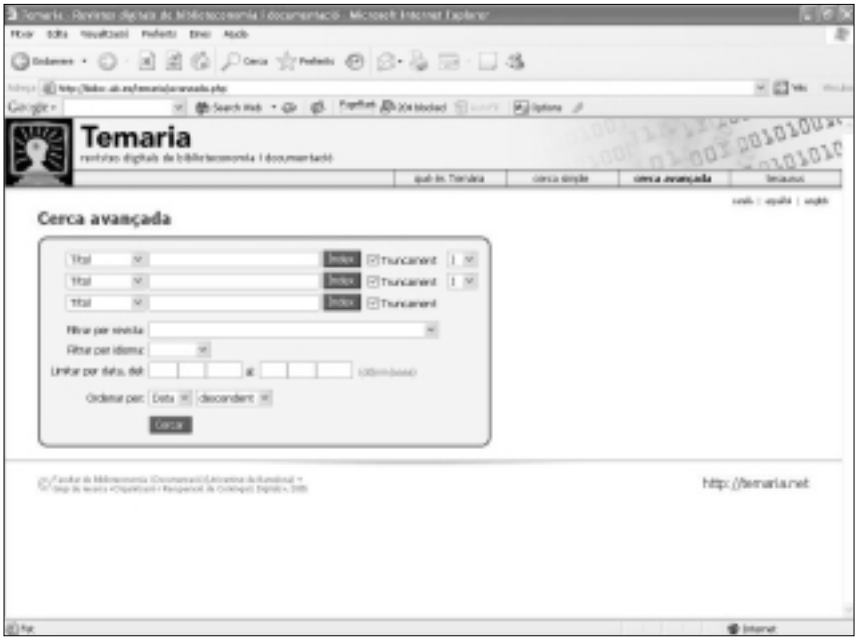

Figura 1. Consulta avançada

tos universitaris de biblioteconomia i documentació, Cuadernos de documentación multimedia i *Hipertext.net*, i ja s'ha arribat a acords amb altres revistes espanyoles (*Item, El profesional de la información*) per incorporar-les progressivament en el portal.

L'objectiu de *Temària* és contribuir a la difusió de les revistes espanyoles de biblioteconomia i documentació, i augmentar-ne la visibilitat. La indexació dels articles de les publicacions que incorpora permetrà explotar al màxim el fons retrospectiu.

Els articles indexats al portal estan escrits, majoritàriament, en català i en castellà, tot i que també se'n poden trobar en anglès. Els usuaris potencials són els professionals, els professors i investigadors universitaris i els estudiants espanyols de l'àmbit de la informació i la documentació que estan interessats a accedir, per mitjà de referències, al text complet de les revistes científiques espanyoles.

El sistema permet que l'usuari pugui consultar els termes del camp temàtic («DC.Subject») en català, castellà o anglès. Aquesta prestació està implementada en la consulta bàsica i en l'avançada.

D'altra banda, també és possible navegar per la jerarquia del tesaurus en aquests tres idiomes. Això permet disposar d'una visió general del camp temàtic, ja que la disposició gràfica en pantalla és molt aclaridora. Soergel (1996) i molts altres investigadors recomanen la presència d'aquesta opció per a la consulta de fons documentals.

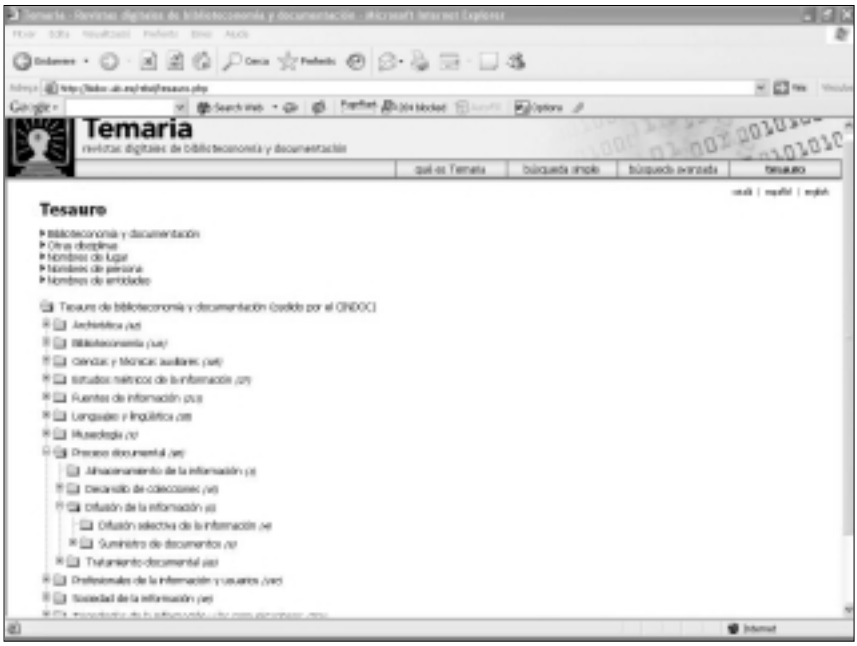

Figura 2. Navegació pel tesaurus

3.1 El tesaurus

L'accés per matèries a la base de dades es duu a terme per mitjà d'un llenguatge postcoordinat d'indexació. Es va valorar l'existència de llenguatges ja elaborats i emprats en l'entorn d'aplicació del portal. De tots ells, el *Tesauro de biblioteconomía y documentación* (en endavant *TBD*) —un llenguatge documental en castellà desenvolupat pel CINDOC— va semblar el més adequat per tres raons principals: la temàtica específica del tesaurus; el grau de coneixement per part dels usuaris potencials del portal, que són també usuaris de les bases de dades del CINDOC, i el fet que el llenguatge disposés d'antuvi de les equivalències dels descriptors en anglès i francès. En el marc del projecte de desenvolupament de *Temària* s'han donat els termes (descriptors i no descriptors) en castellà i anglès, i a més se n'ha afegit la traducció al català.

De tota manera, el *TBD* no pot representar tots els conceptes presents en els continguts indexats en la base de dades, ja que cobreix únicament el camp de la biblioteconomia i documentació. Això vol dir que articles que tracten de la documentació en un camp determinat —documentació mèdica, documentació musical, etc.— no poden representar-se exhaustivament si el *TBD* no es complementa amb un altre tesaurus, de caràcter general, que consideri totes les disciplines. Aquest segon tesaurus ha de servir per representar disciplines altres que la documentació, que ja s'inclou en el *TBD*. Com a base conceptual per al desenvolupament d'aquest segon tesaurus s'ha pres l'*UNESCO thesaurus*. En la base de dades *Temària*, també hi ha la necessitat de representar els noms de lloc, que s'inclouen en l'*UNESCO thesaurus*.

El resultat és un tesaurus general, estructurat en tres parts. Una d'elles, la referida a la documentació, és bàsicament el *TBD*, amb algunes petites variacions (ampliacions de referències, alguns termes nous, etc., que no en modifiquen l'estructura bàsica). La segona, per a la indexació de la resta de disciplines, és un tesaurus general basat en el de la UNESCO, però abreviat i adaptat als continguts que es van trobant: és, en qualsevol cas, un microtesaurus obert, que pot ampliar algun dels seus dominis en funció de les matèries que es vagin incorporant en indexar els articles. La tercera part, estructurada també jeràrquicament, representa la faceta del lloc, ja que ofereix un tesaurus de corònims també basat en l'*UNESCO thesaurus*.

A més, s'hi afegeix una quarta part que agrupa els noms propis, de persones i d'entitats, que s'han fet servir com a matèria. En aquest cas, només es donen les formes autoritzades dels noms en ordre alfabètic, sense una estructuració conceptual.

A la pantalla inicial de cerca per termes del tesaurus, s'ofereixen aquests quatre apartats com a possibilitats per a la cerca, ja dintre d'un d'ells, ja combinant descriptors dels diferents dominis.

3.2 Comentaris

En el tesaurus resultant, s'ha partit d'una estructura jeràrquica construïda en una llengua (el castellà en el cas del *TBD*, el català en el cas de l'adaptació de l'*UNESCO thesaurus* i en la faceta de noms de lloc). Els descriptors s'han traduït a les altres llengües. Es parteix, per tant, d'un tesaurus monolingüe i es tradueix a altres llengües (opció A de construcció de tesaurus multilingüe de la tipologia establerta per Hudon, 1997).

Aquesta opció té l'avantatge que manté la mateixa estructura independentment de la llengua. Aquesta estructura, jeràrquica, es pot assimilar a una classificació: pot ser codificada, «traduïda» a un llenguatge alfanumèric, de manera que cada índex de la classificació tingui un únic codi numèric. És aquest codi, i no els descriptors alfabètics en una llengua concreta, el que s'assigna al registre de la base de dades en l'operació d'indexació d'un article.

Aquesta solució facilita la indexació, ja que només cal assignar un terme d'indexació, el codi, que permetrà tantes formes diferents d'accedir-hi com termes en llengües diferents s'hi associïn. Tanmateix, aquesta solució també planteja problemes difícils de resoldre i que trobem citats recurrentment a la bibliografia sobre tesaurus multilingües (SOERGEL 1996, ap. 2.2). En el nostre cas, podem destacar especialment el fet que les equivalències entre els termes en diferents llengües no sempre són exactes, i això pot implicar que les jerarquies resultants puguin ser diferents.

En general, no es donen problemes entre les equivalències dels termes catalans i espanyols: la proximitat lingüística i cultural és gran, i l'evolució de la terminologia científica en totes dues llengües ha estat similar. En canvi, es poden trobar més problemes amb l'anglès.

Així, una cadena com la següent (no es donen els nivells inferiors), pot ser la mateixa en espanyol i català, ja que la terminologia és equivalent.

Tratamiento documental	Tractament documental
Análisis documental	Anàlisi documental
Análisis de contenido	Anàlisi de contingut
Clasificación	Classificació
Elaboración de resúmenes	Elaboració de resums
Indización	Indexació
Análisis formal	Anàlisi formal
Catalogación	Catalogació
Descripción bibliográfica	Descripció bibliogràfica

QUADRE 1. L'àmbit temàtic del tractament documental (en castellà i en català)

En canvi, en anglès no podem traduir de manera mecànica la cadena anterior: el concepte «anàlisi documental» equival a «cataloging», «classification» pot entendre's com a «classificació i indexació», i només com a «classificació» en un context com el donat per la mateixa estructura. La

distinció entre «anàlisi formal» i «catalogació» és difícil de mantenir en anglès, on totes dues coses són simplement «cataloging» o, especificant, «descriptive cataloging». Si la cadena s'hagués construït pensant ja en una aplicació multilingüe, segurament s'hagués optat per eliminar aquest matís que, fins i tot en català o espanyol, és poc clar, i fondre aquests dos conceptes en un de sol.

Document management
- Cataloging
  - Subject cataloging
    - Classification
    - Abstracting
    - Indexing
  - Descriptive cataloging
    - Bibliographic description

QUADRE 2. L'àmbit temàtic del tractament documental (en anglès)

Quelcom similar es dóna en cadenes amb descriptors que representen conceptes d'una realitat concreta, com poden ser les categories professionals o els graus i centres d'ensenyament, que responen a tradicions culturals diferents i que, per tant, són difícilment equiparables amb l'anglès. Això mateix s'esdevé en els descriptors referits a àmbits de l'administració pública, que varien com a conseqüència d'estructures polítiques o administratives determinades. Així, per exemple, els tipus d'arxius només seran aplicables a la realitat espanyola; si algun article tractés de tipus d'arxius en un altre país, podria ser complicat assignar-hi un terme

Asociacionismo profesional
- Asociaciones profesionales
- Colegios profesionales

Categorías laborales
- Auxiliares
- Ayudantes
- Facultativos
- Personal no profesional

Archivos de la administración
- Archivos autonómicos
- Archivos estatales
- Archivos generales
- Archivos municipales
- Archivos nacionales
- Archivos provinciales
- Archivos regionales

QUADRE 3. Els àmbits temàtics de l'associacionisme i dels arxius de l'administració (en castellà)

equivalent. El mateix passa a l'hora de traduir els descriptors a l'anglès: en primer lloc, no tenim unes administracions equivalents («província», «autonomia», «regió», tenen difícil adaptació en anglès); en segon lloc, els sistemes arxivístics del Regne Unit i dels Estats Units no s'estructuren en aquests nivells.

Fins i tot en la faceta de llocs, on es podria pensar que les equivalències són exactes, hi ha desajustaments que són conseqüència de la diversitat de tradicions culturals. Així, l'*UNESCO thesaurus*, en anglès (i en una traducció literal però inexacta al castellà) dóna el descriptor «Middle East» i, com a equivalent desestimat, «Near East». És, efectivament, la manera habitual d'anomenar en anglès els estats asiàtics dels nostres «Pròxim Orient» i «Orient Mitjà», però no hi ha una coincidència, llevat que es forci el paral·lelisme «Middle East» = «Pròxim Orient i Orient Mitjà». En tot cas, les estructures resultants seran satisfactòries en anglès, però no ho seran ni en castellà ni en català, ja que la nostra tradició hauria donat una estructura que mantindrien separats els dos àmbits geogràfics.

Middle East	**Pròxim Orient**
UF Near East	NT1 Xipre
NT1 Afghanistan	NT1 Israel
NT1 Cyprus	NT1 Jordània
NT1 Gulf States	NT1 Líban
NT2…	NT1 Palestina
NT1 Iran	NT1 Síria
NT1 Iraq	NT1 Turquia
NT1 Israel	
NT1 Jordan	**Orient Mitjà**
NT1 Lebanon	NT1 Afganistan
NT1 Palestine	NT1 Estats del Golf
NT1 Syrian AR	NT2…
NT1 Turkey	NT1 Iemen
NT1 Yemen	NT1 Iran
	NT1 Iraq

Quadre 4. Descriptors geogràfics del Pròxim Orient i de l'Orient Mitjà (en anglès i en català)

Les referències presentaran una problemàtica similar: termes equivalents en una llengua no tindran sentit en una altra, i a l'inrevés.

En casos així, fóra ideal donar cadenes diferents per a cada llengua, de manera que reflectissin fidelment la realitat que rau rere de cadascuna. Serien tres tesaurus diferents que, d'alguna manera, haurien d'estar relacionats entre si, però cadascun amb les seves estructures, les seves relacions, etc. No seria possible un únic codi per a cada concepte del tesaurus i, per tant, ens trobaríem que la indexació s'hauria de fer per triplicat, en cada una de les llengües, creant estructures separades per a la recuperació també de manera separada.

No obstant això, aquesta solució —respectuosa amb la representació del coneixement en cada comunitat cultural— no pot aconseguir la simplicitat en l'aplicació que té la de la traducció: el que és avantatjós en un cas, no ho és en l'altre. La facilitat en l'ús del tesaurus s'ha considerat important i s'ha optat, per tant, per l'opció de la traducció a partir d'una única estructura, solucionant els problemes de manca d'equivalència —que no són tants, potser perquè es tracta d'un àmbit especialitzat del coneixement i on les tradicions terminològiques catalana i espanyola han begut directament de l'anglesa— mitjançant l'ús de termes complexos que reflecteixen el mateix concepte: així, s'han donat com a equivalents «Middle East» i «Pròxim Orient i Orient Mitjà», donant els estats en un únic grup, com en anglès, ja que era l'única manera de mantenir una mateixa estructura en aquest punt. En canvi, en altres casos, s'ha mantingut l'estructura espanyola, ja que s'ha considerat que el gruix de les referències seria sobre casos espanyols i en descriurien la realitat: així passa amb la classe dels tipus d'arxius, que es manté amb la mateixa estructura del *TBD* original en espanyol.

4 Conclusions

Tot i que els documents de la xarxa es troben en idiomes diversos, encara és poc freqüent trobar catàlegs o bases de dades que facilitin l'accés multilingüe per matèries al contingut d'aquests documents. El disseny i desenvolupament de sistemes de recuperació per a recursos web ha de tenir en compte aquesta necessitat, especialment manifesta per als usuaris dels idiomes que no són d'un ús majoritari a la xarxa.

Com s'ha vist, els sistemes que es basen en l'ús de llenguatges documentals poden partir de tesaurus multilingües o optar per cercar alguna forma d'interoperabilitat (mapatge) entre diferents llenguatges documentals. La darrera d'aquestes opcions té un grau més gran de complexitat, ja que implica establir equivalències entre termes de llenguatges documentals d'estructura, de llengua o de graus d'aprofundiment diferents. Això explica que, actualment, es trobin poques experiències en funcionament d'aquest model.

L'altra opció, tot i que comporta algunes dificultats, com el fet que les equivalències entre termes de les diferents llengües del tesaurus no sempre són exactes, sembla la més assequible per facilitar la consulta temàtica multilingüe a col·leccions digitals. Es tracta d'un model que no presenta gaires dificultats d'implantació i que dota els catàlegs i les bases de dades d'una prestació notable amb vista als usuaris. Això és el que s'està duent a terme en el portal *Temària*, que permet accedir al contingut d'articles de revistes espanyoles de biblioteconomia i documentació a partir de metadades DC i d'un tesaurus multilingüe.

Bibliografia

(Adler 2000) Adler, Elhanan. «Multilingual and multiscript subject access: the case of Israel» [recurs electrònic]. En: *Proceedings of the 66th IFLA Council and General Conference, Jerusalem, 13-18 August, 2000*. <www.ifla.org/IV/ifla66/papers/035-130e.htm> [Consulta: 25 novembre 2004].

(Chan 1999) Chan, Lois Mai; Lin, Xia; Zeng, Marcia. «Structural and multilingual approaches to subject access on the web» [recurs electrònic]. En: *Proceedings of the 65th IFLA Council and General Conference, Bangkok, August 20-28, 1999*. <www.ifla.org/IV/ifla65/papers/012-117e.htm> [Consulta: 25 novembre 2004].

(Chan 2002) Chan, Lois Mai; Zeng, Marcia. «Ensuring interoperability among subject vocabularies and knowledge organization schemes: a methodological analysis» [recurs electrònic]. En: *Proceedings of the 68th IFLA Council and General Conference, Glasgow, August 18-24, 2002*. <www.ifla.org/IV/ifla68/papers/008-122e.pdf> [Consulta: 25 novembre 2004].

(Clavel-Merrin 1999) Clavel-Merrin, Genevieve. «The need for co-operation in creating and maintaining multilingual subject authority files» [recurs electrònic]. En: *Proceedings of the 65th IFLA Council and General Conference, Bangkok, August 20-28, 1999*. <www.ifla.org/IV/ifla65/papers/080-155e.htm> [Consulta: 25 novembre 2004].

(Clavel-Merrin 2004) Clavel-Merrin, Genevieve. «MACS (Multilingual access to subjects): a virtual authority file across languages». *Cataloging and classification quarterly*, vol. 39, no. 1-2 (2004), p. 322-330. Versió electrònica disponible a: <http://eprints.rclis.org/archive/00000277/01/clavel-merrin_eng.pdf> [Consulta: 4 abril 2005].

(Davies 2003) Davies, Ron. «Models for multilingual subject access in online library catalogues: the ILO experience» [recurs electrònic]. En: *ELAG 2003: Cross Language Applications and the Web. 27th Library Systems Seminar, Bern (Switzerland), 2-4 April 2003*. <www.elag2003.ch/pres/pres_davies.pdf> [Consulta: 25 novembre 2004].

(Doerr 2001) Doerr, Martin. «Semantic problems of thesaurus mapping» [recurs electrònic]. *Journal of digital information*, vol. 1, issue 8, article no. 52 (2001-03-26). <http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Doerr/> [Consulta: 25 novembre 2004].

(Fátima 2004) Fátima Pellizzon, Rosely de. «Pesquisa na área de saúde: 1 – base de dados DeCS (Descritores em Ciencias da Saúde)» [recurs electrònic]. *Acta cirurgica brasileira*, v. 19, n. 2 (mar.-abr. 2004). <www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-86502004000200013&lng=pt&nrm=iso> [Consulta: 6 abril 2005].

(Francu 1996) Francu, Victoria (1996). «Building a multilingual thesaurus based on UDC». En: ISKO. Conference (1996: Washington). *Knowledge organization and change: proceedings of the 4th International ISKO Conference, Washington, 15-18 July, 1996* (Frankfurt/Main: Indeks, 1996), p. 144-154.

(Hudon 1997) Hudon, Michele. «Multilingual thesaurus construction-integrating the views of different culture in one gateway to knowledge and concepts». *Knowledge organization*, 24, no. 2 (1997), p. 84-91.

(ILO 1998) Organización Internacional del Trabajo. *ILO thesaurus: labour, employment, and training terminology*. 5th ed. Geneva, Switzerland: ILO, 1998.

(Keränen) Keränen, Susanna. «Cultural and linguistic differences in digital storage and retrieval of information» [recurs electrònic]. <www.abo.fi/instut/diginfo/plan.html> [Consulta: 25 maig 2004].

(LANDRY 2000) LANDRY, Patrice. «The MACS project: multilingual access to subjects (LCSH, RAMEAU, SWD)» [recurs electrònic]. En: *Proceedings of the 66th IFLA Council and General Conference, Jerusalem, August 13-18, 2000*. <www.ifla.org/IV/ifla66/papers/165-181e.pdf> [Consulta: 25 novembre 2004].

(LANDRY 2003) LANDRY, Patrice. «MACS update: moving toward a Link Management Production Database» [recurs electrònic]. En: *ELAG 2003: Cross Language Applications and the Web. 27th Library Systems Seminar, Bern (Switzerland), 2-4 April 2003*. <www.elag2003.ch/papers/MACS-ELAG-article.pdf> [Consulta: 6 abril 2005].

(MATTHEWS 2000) MATTHEWS, Brian; WILSON, Michael. «Multilingual metadata to access social science data». *ERCIM news*, no. 41 (April 2000). <www.ercim.org/publication/Ercim_News/enw41/matthews1.html> [Consulta: 6 abril 2005].

(MOCHÓN 2002) MOCHÓN, Gonzalo; SORLI, Ángela. *Tesauro de biblioteconomía y documentación* [recurs electrònic]. Madrid: Consejo Superior de Investigaciones Científicas, 2002. <http://pci204.cindoc.csic.es/TESAUROS/Bib_Doc/Bib_Doc.htm> [Consulta: 6 abril 2005].

(NICHOLSON 2002) NICHOLSON, D. «Subject-based interoperability: issues from the High Level Thesaurus (HILT) project» [recurs electrònic]. *Proceedings of the 68th IFLA Council and General Conference, Glasgow, August 18–24, 2002*. <www.ifla.org/IV/ifla68/papers/006-122e.pdf> [Consulta: 6 abril 2005].

(OARD 1996) OARD, Douglas W.; DORR, Bonnie J. «A survey of multilingual text retrieval» [recurs electrònic]. 1996. <http://citeseer.ist.psu.edu> [Consulta: 25/11/2004].

(OARD 1997) OARD, Douglas W. «Cross-language information retrieval bibliography» [recurs electrònic]. 1997. <http://citeseer.ist.psu.edu> [Consulta: 25 novembre 2004].

(OARD 1999) OARD, Douglas W.; RESNIK, Philip. «Support for interactive document selection in cross-language information retrieval». *Information processing & management*, no. 35 (1999), p. 363-379.

(OGDEN 2000) OGDEN, William C.; DAVIS, Mark W. «Improving cross-language text retrieval with human interactions» [recurs electrònic]. En: *Proceedings of the 33rd Hawaii International Conference on System Sciences, January 2000*. <http://crl.nmsu.edu/Research/Projects/tipster/ursa/Papers/Hawaii.pdf> [Consulta: 9/05/2005].

(PAVANI 2001) PAVANI, Ana M. B. «A model of multilingual digital library». *Ciência da Informação*, v. 30, n. 3 (set.-dez. 2001), p. 73-81. Versió electrónica disponible a: <www.scielo.br/pdf/ci/v30n3/7289.pdf> [Consulta: 9 maig 2005].

(PETERS 1997) PETERS, Carol; PICCHI, Eugenio «Across languages, across cultures: issues in multilinguality and digital libraries» [recurs electrònic]. *D-Lib*, (May 1997). <www.dlib.org/dlib/may97/peters/05peters.html> [Consulta: 25 novembre 2004].

(RODRÍGUEZ 1998) RODRÍGUEZ CAMIÑO, Reinaldo. «*MeSH* o *DeCS*: algunas consideraciones sobre la indización biomédica» [recurs electrònic]. *ACIMED*, vol. 6, núm. 3 (1998), p. 163-170. Versió electrònica disponible a: <http://eprints.rclis.org/archive/00002037/01/aci04398.pdf> [Consulta: 4 abril 2005].

(SAVOY 2003) SAVOY, Jacques. «Cross-language information retrieval: experiments base on CLEF 2000 corpora». *Information processing & management*, no. 39 (2003), p. 75-115.

(SOERGEL 1996) SOERGEL, Dagobert. «Multilingual thesauri in cross-language text and speech retrieval» [recurs electrònic]. En: *Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, Stanford University, March 24-26, 1997*. <www.ee.umd.edu/medlab/filter/sss/papers?soergel.ps> [Consulta: 4 abril 2005].

(SOERGEL 2004) SOERGEL, Dagobert et al. «Reengineering thesauri for new applications: the AGROVOC example» [recurs electrònic]. *Journal of digital information*, vol. 4, issue 4, article no. 257 (2004-03-17). <http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Soergel/> [Consulta: 4 abril 2005].

(ZENG 2004) ZENG, Marcia; CHAN, Lois Mai. «Trends and issues in establishing interoperability among knowledge organization systems». *Journal of the ASIS*, vol. 55, no. 5 (2004), p. 377-395.