Número 9, desembre 2002

Avaluació de les metadades com a eina d'indexació i recuperació al web de la Biblioteca de la Universitat de Barcelona

SÍLVIA ARGUDO PLANS[1]
Facultat de Biblioteconomia i Documentació
Universitat de Barcelona
argudo@fbd.ub.es

Resum

Coincidint amb la renovació del web de la Biblioteca de la Universitat de Barcelona, al mes de juny s'inicià un procés d'avaluació amb l'objectiu de replantejar la utilització de les metadades com a eina d'indexació i recuperació de continguts de pàgines web. L'article s'estructura en tres parts principals. En una primera part, s'estableixen els antecedents del web de la Biblioteca, l'origen i la forma de la inclusió de metadades en les seves pàgines i l'evolució del tema fins a arribar al moment de l'avaluació. Tot seguit, s'exposen totes les dades analitzades en el procés d'avaluació fet, partint de conceptes i aspectes teòrics fonamentals de l'anàlisi de contingut. L'apartat de conclusions amb què finalitza l'article parteix de la interpretació de les dades dels dos processos de manera relacionada per oferir un seguit d'indicacions o pautes que cal tenir en compte en el replantejament de la utilització de metadades com a eina de representació de contingut i de recuperació d'informació de les pàgines web de la BUB.

1 Introducció

Les experiències amb metadades de les quals ens arriba notícia en forma de publicacions s'emmarquen sempre en projectes molt concrets. Amb poques excepcions, es tracta de projectes planificats prèviament de manera més o menys detallada, que abasten recursos especialitzats temàticament, que utilitzen un conjunt de metadades específic creat o adaptat com a part del mateix projecte i amb un sistema de recuperació dissenyat —o configurat també específicament— per al conjunt de recursos concret que s'hi inclou.

Poques vegades es descriuen experiències fetes en entorns genèrics. No se solen estudiar els resultats de l'ús de metadades en pàgines de llocs web genèrics i pluridisciplinaris, ni especialitzats en un tema concret ni amb recursos d'una tipologia documental concreta. En els pocs casos en què això es dóna,[2] tampoc no es fa una anàlisi àmplia, que abasti tant els aspectes d'indexació de les pàgines com el procés de la recuperació i els resultats obtinguts per l'usuari en aquesta fase.

El fet de tractar-se d'un web genèric i no d'un conjunt de recursos especialitzats en una matèria, fa que els resultats de l'anàlisi siguin forçosament diferents d'altres estudis publicats, sobretot quan s'afronta l'estudi de la recuperació. Els usuaris són variats, no estem davant d'usuaris experts en una matèria que busquen recursos especialitzats en el tema, que en coneixen el vocabulari i que estan acostumats a utilitzar sistemes de recuperació d'informació com poden ser bases de dades especialitzades. Entre el ventall d'usuaris d'un web genèric trobem diferents nivells cognitius, diferents necessitats d'informació i diferents maneres d'expressar-les. En els escassos treballs existents sobre les característiques de la recuperació d'informació al web, ja s'apunten les diferències trobades entre aquesta recuperació en entorns genèrics amb usuaris diversos i la que es dóna en entorns especialitzats amb usuaris més entrenats.[3]

La indexació i la recuperació estan estretament relacionades, de manera que el que es fa en el

procés d'indexació determina gran part del que passarà després en la fase de recuperació. A la inversa, el que passa durant la recuperació d'informació, hauria de determinar les decisions i polítiques que afecten la indexació. Es pot dir que hi ha un consens general respecte d'aquestes afirmacions i no sembla probable que ningú estigui en desacord pel que fa a la formulació teòrica. Tanmateix, no sempre es té en compte a efectes pràctics; poques vegades en el dia a dia professional es pot disposar de temps per replantejar el tractament i la representació del contingut dels documents a partir de l'anàlisi del que passa en el procés de recuperació que duu a terme l'usuari.

Aquest és l'objectiu que ha estat present durant l'avaluació del sistema de metadades de la Biblioteca de la Universitat de Barcelona (BUB). L'anàlisi conjunta de la indexació i de la recuperació, l'estudi de les dades en ambdós processos a partir de la seva estreta i indiscutible relació, hauria de donar les pautes per al replantejament planificat de la utilització de metadades a les pàgines web de la Biblioteca.

2 Antecedents: per què, quan i com

2.1 El web de la BUB

El 1995 van aparèixer les primeres pàgines web de la Biblioteca de la UB. El setembre de 1996 es designà una persona que se n'havia de fer càrrec, coordinant un equip de gent, planificant i organitzant la tasca, com una més de les tasques bibliotecàries. En aquell moment, la presència al web d'institucions acadèmiques i bibliotecàries ja es considerava de prou importància per replantejar-se el tema de manera seriosa. Especialment en l'àmbit bibliotecari, ja s'havien vist les possibilitats de la nova eina com a element especialment valuós per a la difusió i l'accés a la informació.

El web s'havia utilitzat en un principi per oferir informació sobre les institucions, reproduint els tríptics i guies que ja existien en paper. Hi ha un moment, entre el 1996 i el 1997, en el qual es començà a anar més enllà i es pensà a aprofitar les capacitats de la nova tecnologia per oferir serveis d'informació, donar accés als recursos propis d'informació de la Biblioteca, atès que ja es disposava de recursos en format electrònic que permetien aquest pas endavant. Aquest canvi d'enfocament en la utilització del web es va donar de manera més o menys simultània arreu.

2.2 Per què

Amb aquesta nova utilització del web, el volum de pàgines de la BUB, com el de la resta d'institucions, començà a fer-se prou gran per pensar que, amb la navegació entre pàgines que proporcionaven els enllaços, no n'hi havia prou. Calia disposar d'un sistema de recuperació d'informació, un sistema que permetés fer cerques i recuperar la informació de manera ràpida i efectiva.

Aleshores, ja havien aparegut els grans cercadors, sistemes que permetien introduir termes de cerca i recuperar pàgines de diversos webs que, en principi, tractaven del tema buscat. Aquests sistemes feien (i fan) les cerques comparant els termes introduïts per l'usuari en una casella de cerca amb els termes que apareixien a les pàgines web que tenien indexades. Es tractava de cerques fetes en el text complet de les pàgines, tot i que en realitat només es feien en una part del text (primers paràgrafs, primers 300 caràcters, etc., depenent de cada cercador). Com tots sabem, la presència d'un terme en un document no assegura que el document tracti d'aquest tema. Si no s'especifica res més, el terme *Cela* pot donar com a resultat pàgines web que tractin de la persona *Cela* (matèria) o pàgines que reprodueixin obres de *Cela* (autor). Bàsicament per aquest motiu, els cercadors donaven (i donen) un resultat amb molt soroll, massa pàgines no rellevants i no pertinents, que no feien cap servei a l'usuari. Una de les possibles solucions al problema es materialitzà amb la proposta dels quinze elements bàsics Dublin Core en una reunió celebrada a Dublin (Ohio) el 1995 sota els auspicis de l'OCLC. Es tractava d'un sistema de catalogació, de descripció de les pàgines web, a partir de quinze elements, o camps d'informació, anomenats *metadades*, que s'havien d'incloure a les pàgines i que permetrien recuperar-les de manera pertinent. La idea adquirí importància i ressò progressius des d'un principi. Alguns grans cercadors començaren a incorporar la capacitat de fer cerques en els camps de metadades per tal de millorar els resultats disminuint el soroll.[4] Van aparèixer ràpidament programes o petits motors de cerca gratuïts i disponibles a la xarxa Internet, que podien fer cerques tenint en compte aquestes metadades i ordenar els resultats per rellevància segons si el terme s'havia localitzat en un camp de metadades o en la resta del text complet de les pàgines. Aquests programes es podien descarregar de la xarxa per instal·lar-los i utilitzar-los en webs de manera local; i oferien, a més a més, la possibilitat de consultar els fitxers Log, o registres de tot el que fan els usuaris durant el procés de recuperació, de manera que es podia saber què havien buscat, com ho havien fet i quin resultat havien obtingut. Ara bé, aquests programes i els cercadors generals existents no assumien, ni ho fan actualment, el format de metadades de la proposta Dublin Core, sinó que utilitzen un format genèric que respon, amb tot, a la mateixa idea. Així, per exemple, en

lloc de buscar els termes en un camp anomenat “DC.Description”, ho fan en el camp “Description”.

2.3 Quan i com

El 18 de setembre de 1997 se celebrà una reunió de l'equip de persones responsables del web de la Biblioteca de la UB amb l'objectiu de tractar la incorporació de metadades a les pàgines. Tot i que ja feia un parell d'anys de la proposta Dublin Core, aquest tema tot just s'havia començat a tractar en l'entorn bibliotecari espanyol i català, i el volum de publicació sobre metadades tampoc no era encara gaire important.

Una cerca feta sobre això a diverses bases de dades mostra la diferència en el volum de publicació abans i després de 1997.

Taula 1. Resultats de la cerca: metadata OR Dublin Core OR metadatos OR metadades

	31/12/1997	**01/01/1998**
Library Literature & Information Science Full Text	40	217
LISA: Library and Information Science Abstracts	70	390
ERIC	32	194
DATATHÉKE	18	71
CSIC	1	23
BEDOC	0	7

En l'esmentada reunió, es proporcionà als assistents informació sobre la proposta Dublin Core, s'explicà en què consistia i es facilitaren exemples de la inclusió de metadades a pàgines web. A més a més, s'anuncià que ja es disposava d'un programa capaç de fer cerques a partir d'aquests nous elements, instal·lat i provat amb èxit,[5] i es donaren unes pautes molt bàsiques i genèriques per començar a treballar.

Els criteris facilitats eren els següents:

- El format de metadades que la BUB adoptaria seria el genèric. Malgrat que el programa de cerca local es pogués modificar per tal d'entendre el format Dublin Core, s'entenia que el model genèric feia possible que els grans cercadors més importants del moment aprofitessin les metadades, amb la qual cosa es facilitava la difusió de pàgines web de la Biblioteca.
- En principi, les metadades que s'utilitzarien serien tres: “Author”, “Description” i “Keywords” (a més a més, es recordà l'obligatorietat de fer servir l'etiqueta “Title” en la capçalera dels documents, utilitzada també pels cercadors com a font d'informació i presentació de resultats). Els motius d'aquesta decisió eren senzills: les pàgines web de la Biblioteca no contenien informació que pogués interessar recuperar per la seva data de publicació, pel seu format, o per la resta d'informació susceptible de ser representada amb metadades. Aquesta informació, d'altra banda, tampoc no es feia constar en les metadades utilitzades habitualment pels cercadors generals en fer les cerques. Fins i tot, el camp “Author” era una dada que interessava només en l'àmbit intern, amb vista a la gestió administrativa del web i, per tant, aquesta metadada no s'utilitzaria per a les cerques dels usuaris.[6]
- El camp o metadada “Keywords” s'ompliria amb paraules clau que servissin per expressar el contingut de la pàgina web. Al camp “Description” es faria un breu resum del contingut en una o dues frases que havien d'imitar al màxim la forma natural d'expressar-se i d'utilització del llenguatge de qualsevol persona. D'aquesta manera, els dos camps esdevenien complementaris i oferien dues maneres diferents d'accedir a la informació quant a la forma.
- El programa de recuperació faria la cerca al text complet dels documents, incloses les metadades, però, a l'hora de presentar els resultats, els ordenaria per rellevància. El programa assignava una puntuació inicial —un 1— a tots els documents trobats i, a l'hora d'ordenar els resultats per presentar-los a l'usuari, multiplicava aquesta puntuació inicial per 4 si trobava el terme a la metadada “Keywords”, per 2 si el trobava a “Description”, i per 1 si el trobava a la resta del document.

No es va donar cap altra pauta, no es va parlar de fer servir cap vocabulari controlat, de singulars o plurals, de termes compostos i termes simples, etc. Era el mínim que es va considerar necessari per començar i ja es veuria què passava quan es fes una primera avaluació.

De tant en tant, s'examinaven els registres de les cerques per veure què estava passant i es

veien algunes qüestions preocupants, si més no curioses, com ara que l'usuari feia cerques que semblaven més pròpies per cercar documents en un catàleg bibliogràfic que en un conjunt de pàgines web. També es revisaven les metadades assignades pel personal responsable de les diferents pàgines i se'n feien alguns comentaris. Mai no es va trobar el moment i el temps de fer cap avaluació seriosa, ja que ben aviat es van produir canvis.

El 1998, hi va haver un canvi de cercador i es passà a utilitzar el programa *Search'97* de Verity, adquirit per la UB. Es treballà amb un becari responsable del programa al Centre d'Informàtica per tal de configurar les característiques de les cerques a la col·lecció de pàgines de la Biblioteca. Aquestes cerques serien les úniques de tot el web de la UB que utilitzarien les metadades, ja que no es feien servir en la resta de pàgines de la institució. La cerca es feia per defecte només a les metadades, però hi havia una opció amb què l'usuari podia seleccionar de fer-la a text complet. Es podien utilitzar operadors booleans, buscar expressions o frases, fer truncaments, etc. Aquestes capacitats, sumades als avantatges del nou programa quant a potència i rapidesa en la cerca, van fer decidir el canvi de cercador. Malgrat que la qualitat dels resultats de l'antic programa era bona, començava a plantejar problemes de lentitud i no tenia possibilitats d'utilitzar operadors.

Durant uns anys no es disposà de fitxers Log,[7] no es podia veure què passava amb la recuperació, ni quin era el resultat de la utilització de les metadades. A més a més, com passa a tot arreu, la feina urgent sempre va retardant la feina important, de manera que no va ser fins el moment de la reorganització total del web de la Biblioteca, el gener de 2002, que es plantejà la necessitat de fer una revisió a fons de la qüestió de les metadades.

3 El procés d'avaluació

3.1 Inici de l'avaluació

En el projecte de reorganització del web de la Biblioteca que s'elabora durant l'any 2001, queden incloses les metadades com a element imprescindible a tenir en compte. L'enfocament bàsic del nou web és l'organització temàtica dels recursos d'informació i les metadades es consideren una eina bàsica que complementa la navegació per l'estructura visible del web a l'hora de facilitar la recuperació dels recursos als usuaris. Com que es fan canvis importants de la disposició dels elements, és important disposar d'una eina que permeti trobar ràpidament allò que ja no està al mateix lloc de sempre; es tracta de compensar d'alguna manera la desorientació inicial de l'usuari davant dels canvis. A més a més, el volum actual de pàgines del web, més de 2.900, fa necessari disposar d'un bon sistema de recuperació.

Això coincideix amb una nova situació pel que fa al cercador. La Universitat ha decidit tornar a concedir al sistema de recuperació la importància que es mereix, coincidint amb un projecte de reestructuració del web de la UB, per la qual cosa hi destinarà personal i es treballarà en el tema.

Hi ha un moment en què es donen dues situacions que fan possible la revisió de les metadades i de l'ús que se'n fa per a la recuperació: d'una banda, ja es pot disposar dels registres de cerques dels usuaris a la col·lecció de pàgines de la Biblioteca, atès que el Centre d'Informàtica els pot facilitar; d'altra banda, el nou web es considera ja implementat i en funcionament, de manera que el volum de feina urgent disminueix de manera important i permet dedicar-se a diverses qüestions primordials pendents d'abordar.

3.2 Objectius de l'avaluació

L'objectiu general de l'avaluació és replantejar tot el sistema d'assignació de metadades a partir de l'establiment de pautes i criteris per a la indexació, i de l'elaboració d'eines d'ajuda.

En l'avaluació es vol veure com s'està fent la feina des del punt de vista de la indexació, com s'està utilitzant des del punt de vista de la recuperació i quins resultats dóna. L'anàlisi conjunta d'aquestes qüestions hauria d'oferir uns resultats que servissin com a guia i orientació a l'hora d'establir els criteris i les pautes necessaris.

Els objectius específics de l'avaluació són:

- Elaborar una llista de descriptors no especialitzats que pugui ser consultada pel personal responsable de l'assignació de metadades. En una fase posterior, aquesta llista haurà de completar-se amb diverses eines d'ajuda, algunes de les quals poden ser proporcionades pel mateix cercador, com poden ser un sistema d'ajuda per al control de la sinonímia, mecanismes àgils d'actualització, consulta i descàrrega de la llista, etc.
- Establir equivalències en castellà i anglès per a les paraules clau en català per tal de facilitar la cerca en aquests idiomes, atès l'alt i creixent percentatge d'estudiants sense coneixements de català presents a la UB i el nombre de consultes externes a l'àmbit català

que es detecta en les estadístiques de visites al web de la BUB.
- Establir una sèrie de pautes i criteris conceptuals que s'apliquin a l'hora d'assignar el contingut de les metadades.
- Establir una sèrie de pautes i criteris referits a la forma dels descriptors que s'apliquin a l'hora d'afegir nous descriptors.

3.3 Metodologia d'avaluació

Pel que fa la metodologia seguida, com s'ha comentat més amunt, la revisió consisteix en dos aspectes que caldrà relacionar: l'anàlisi de la indexació i l'anàlisi de la recuperació. Per a l'estudi de cadascun d'aquests aspectes es procedeix de manera diferent, ja que cal analitzar dades diferents.

Per revisar la indexació, es demana al personal responsable del web que trameti a la Unitat de Tecnologies de la Informació un fitxer amb tots els termes presents a la etiqueta “Keywords” de les seves pàgines. Després, aquests termes són inclosos conjuntament en un fitxer d'*Excel* per examinar-los i analitzar-los detalladament. Per estudiar aspectes qualitatius de la indexació, cal examinar les pàgines web a les quals s'assignen determinats termes.

Per a l'estudi de la recuperació es fa servir l'anomenada anàlisi transaccional o anàlisi de fitxers Log. Es demanen al Centre d'Informàtica els fitxers amb els registres de les cerques dels usuaris corresponents als mesos de maig i juny. La tria d'aquests mesos concrets respon principalment al fet que són els dos últims de què es disposa i per això es considera que el nou web ja està prou assentat perquè els usuaris en coneguin l'estructura. Si es trien mesos anteriors, el volum de cerques es pot veure alterat per la nova disposició de la informació, que fa que els usuaris no la trobin fàcilment amb la simple navegació i utilitzin el cercador amb més freqüència de l'habitual. D'altra banda, al mes de juny no hi ha classes i això fa que les cerques siguin diferents que en altres mesos com ara maig o abril. Agafant maig i juny, s'abasta la diferent tipologia de cerca d'informació, la corresponent a un període normal de docència i la corresponent a un període d'exàmens. Com que no es disposa de cap programa d'anàlisi de fitxers Log, les dades es tracten també amb el programa *Excel* pel que fa a l'estudi quantitatiu. Els aspectes qualitatius s'examinen manualment utilitzant mostres formades pels conjunts de cerques en nombre prou representatiu. Finalment, s'extreuen unes conclusions amb el resum del que s'ha trobat i les qüestions que cal tenir en compte per a la planificació efectiva del sistema de metadades.

3.4 Anàlisi de la indexació

Una vegada ordenats els termes alfabèticament i eliminats els duplicats, es comptabilitzen 815 paraules clau diferents assignades a una mica més de 2.900 documents.[8]

El nombre de duplicats no és gaire elevat i es refereix gairebé de manera exclusiva a termes genèrics i/o que designen serveis de la biblioteca, com ara *préstec*, *préstec interbibliotecari*, *revistes*, *horaris*, etc. El nivell baix de duplicitat era esperable atès que, a diferència del que passa en un catàleg, on és normal que hi hagi documents que tractin d'un mateix tema, en un únic lloc web com és el de la Biblioteca no és freqüent que hi hagi gaires pàgines amb un contingut similar.

3.4.1 Forma i categoria gramatical

Ús de singulars o plurals: la majoria de les paraules clau, amb poques excepcions, segueixen, sense haver-s'ho proposat, les recomanacions sobre el tema que es troben a la norma de creació de tesaurus monolingües, UNE 50 106:1990 (Asociación Española de Normalización y Certificación, 1999, p. 87–88). És a dir, per representar entitats concretes i quantificables es fa servir el plural i per designar entitats concretes incomptables o conceptes abstractes —com ara propietats, fenòmens, disciplines, etc.— la forma utilitzada és el singular.

Respecte de les categories gramaticals, la majoria de termes són substantius, solts o adjectivats, tot i que també s'han trobat adjectius solts (*gastrointestinal*, *digital*, *espanyola*) i alguns verbs en infinitiu (per exemple *citar*, *trobar*).

3.4.2 Sinònims i formes variants

La quantitat de formes variants i sinònims que s'han trobat és realment alta, però es refereix de manera gairebé exclusiva a termes genèrics, no als especialitzats temàticament. La majoria de responsables de pàgines web procura pensar en les equivalències més comunes i conegudes a l'hora d'assignar les metadades i les fa constar a la mateixa pàgina, però sempre en falta alguna

en la qual no es pensa o es pensa més tard i no es recuperen pàgines antigues per afegir el nou descriptor.

Cal identificar i agrupar totes les equivalències i afegir-ne d'altres, també les corresponents a les llengües castellana i anglesa. També és necessari treballar en la normalització de noms propis, del quals s'utilitzen diverses formes.

Amb tot, aquest punt era un dels que ja se sabia amb certesa que necessitava alguna actuació. És un dels problemes bàsics de les llistes de descriptors lliures per oposició als llenguatges controlats. Un objectiu posterior que ja s'ha declarat més amunt, és incloure al sistema de cerca un diccionari de termes, de manera que el programa identifiqui les equivalències i així no calgui haver-hi de pensar i haver-les de posarcada vegada.

Un aspecte afegit que s'ha detectat en aquest punt és la necessitat d'explicar amb més detall el funcionament del cercador a tot el personal. Moltes de les formes variants que s'han afegit, contràriament al que ha pensat l'indexador en el moment de fer-ho, no suposen un augment de les possibilitats de recuperació. Per exemple, no té cap sentit fer constar les diferents formes *Companys*, *Lluís Companys* i *Companys, Lluís* tenint en compte:

a) que el cercador busca en principi i per defecte cadenes de caràcters i

b) que el cercador entén les comes entre termes com si fossin l'operador booleà AND.

És a dir, triant la forma *Lluís Companys* per indexar, es recuperarà el document tant si l'usuari busca la forma de nom i cognom sencers, com si busca només el cognom, com si busca la forma invertida.

3.4.3 Conceptes simples i conceptes compostos

Quan es parla d'aquest tema, sempre apareixen dues qüestions diferents tractades conjuntament. Provarem de tractar aquestes dues qüestions per separat, tot i que no sempre és senzill de fer-ho de manera clara i entenedora. D'una banda, hi ha termes simples i compostos, segons si estan formats per una sola paraula —o unitat lingüística— o per més d'una; l'altra qüestió és si aquests termes representen conceptes simples o bé compostos. En un llenguatge postcoordinat, com són les llistes de descriptors lliures o els tesaurus, la recomanació general, tal com apareix a la norma ja esmentada de construcció de tesaurus, és reduir al màxim la representació de conceptes compostos. L'usuari ja farà la combinació amb operadors booleans en la fase de recuperació per tal de recuperar conceptes compostos o complexos. Tanmateix, això no sempre és tan senzill i la mateixa norma, després de recordar aquesta regla general de simplificació i utilització de termes simples, ofereix unes pautes per ajudar a decidir en quins casos cal fragmentar sintàcticament els termes compostos i en quins casos és preferible no fer-ho.

Examinant les metadades de la BUB es veu que, del total de 815 termes, 260 són compostos i 555 simples.

D'aquests 555 termes simples, aproximadament el 73% corresponen a conceptes simples representatius del contingut de pàgines web sense necessitat de combinar-los amb altres.[9] Per exemple es pot esmentar *acrònims*, *educació* o *jurisprudència*.

En pocs casos, el 4%, es troben termes simples que designen conceptes compostos, com és el cas dels termes *aromateràpia* o *bioètica*. La fragmentació hauria de ser en aquests casos semàntica i no resultaria de gaire utilitat en la recuperació, ja que es tracta de termes molt establerts en el camp de la seva especialitat temàtica. Un 23% dels termes simples no tenen cap sentit si no és per combinar-los amb altres. És a dir, tot i que els termes representen conceptes existents en el món real, no hi ha cap pàgina en el web de la Biblioteca que tracti dels conceptes que aquests termes representen. Per exemple, hi ha termes com ara *codis*, *instruccions* o *organismes*, però no hi ha pàgines que parlin de codis, instruccions o organismes en general, sinó que parlen específicament de *codis d'accés a la intranet*, *instruccions d'instal·lació de programes*, *organismes públics*, *organismes internacionals*, etc.

Si considerem les pautes que estableix la norma, la majoria d'aquests termes simples serien correctes, ja que les expressions es poden fragmentar. Tot i oferir certes pautes, la norma no soluciona tots els problemes. Un exemple: cal fragmentar la forma *acidesa del sòl* (*acidesa* + *sòl*) però no es pot fragmentar *sòls àcids*; el que no diu és quina de les dues formes resulta més adequada emprar. Des d'un punt de vista pràctic, doncs, arriba un moment en què l'únic criteri que pot ajudar és el que s'obté de respondre la següent pregunta: es correspon la fragmentació o no fragmentació d'aquest tipus de termes amb la realitat de la recuperació? Tot i estar parlant d'un sistema postcoordinat, no sembla probable que cap usuari provi de recuperar informació sobre *codis d'accés a la intranet* combinant els termes *codis*, *accés* i *intranet*. La hipòtesi que

plantegem és que l'usuari fa les cerques amb llenguatge natural, tal com parla, qüestió que, tanmateix, comprovarem en l'anàlisi de la recuperació.

Si analitzem el terme *instruccions* pensant en les pàgines del web que el contenen com a descriptor, veiem que el problema no es refereix només a la intenció de pre- o postcoordinació de conceptes o a la fragmentació de termes compostos, ni tan sols a qüestions d'especificitat, de la qual parlarem més endavant, sinó a la necessitat de representar la forma, a més de la matèria. Per exemple, si el tema de què tracten les pàgines és la instal·lació de programes client, les instruccions en aquest cas es poden considerar la forma que pren la matèria. El tema està tractat des del punt de vista de donar instruccions, es volen donar instruccions, són instruccions, però la matèria no són les instruccions. S'ha de representar la forma amb descriptors igual que les matèries? L'usuari busca forma o només matèries? És una altra qüestió que serà necessari aclarir, tot i que en aquest exemple concret de les instruccions sembla bastant probable que la forma es correspon amb l'expressió d'una necessitat de recuperació.

Altres termes simples trobats, com per exemple la paraula *anys*, encara són més dubtosament útils com a descriptors. Aquesta paraula clau està assignada a una pàgina on s'explica com interpretar les pantalles de fons de les revistes <http://www.bib.ub.es/bub/guiahis.htm>, juntament amb els descriptors *ajuda*, *trobar*, *revistes*, *volums*, *números*, *històrics*, i *catàleg*.

Tot i que la quantitat de termes simples sense sentit, pensats només per combinar amb molts altres, és petita, un 23%, caldrà establir alguns criteris que ajudin a donar coherència a la indexació en aquest aspecte.

Els termes compostos que trobem a la llista de descriptors representen gairebé sempre una intenció d'especificació, concreció o aclariment de matèries més genèriques. Alguns exemples són: *cartes nàutiques*, *infermeria pediàtrica* o *literatura juvenil*, que s'utilitzen en lloc dels termes *infermeria* + *pediatria*, *literatura* + *joves*.

En l'entorn acadèmic en què es mouen els usuaris de la UB (i probablement també fora d'aquest entorn), termes com ara *literatura juvenil* o *infermeria pediàtrica* formen part del llenguatge habitual, són d'ús comú. No és probable que un estudiant d'infermeria faci una cerca combinant *infermeria* i *pediatria*; el més probable és que la faci per la forma precoordinada, pel terme compost directament. Cal tenir en compte, a més a més, que aquest tipus de precoordinació adjectivant un substantiu assegura en la majoria de casos la representació exacta i específica d'un concepte, ja que, per exemple, *literatura juvenil* es refereix a literatura escrita per als joves com a destinataris, mentre que la combinació *literatura* + *joves* pot referir-se a literatura escrita per gent jove, destinada a joves i també que parli sobre joves.

La realitat de les metadades de la BUB és que, tot i tractar-se d'un llenguatge lliure postcoordinat, on hi hauria d'haver, i de fet hi ha, majoria de termes simples, s'observa una presència important de termes compostos que representen conceptes compostos.

Malgrat la decisió final que es pugui prendre al respecte, és important fer notar que la manca de pautes i la impossibilitat de consultar la llista de descriptors existents provoquen fets com el de l'existència simultània dels següents termes: *premis Nobel*, *economia*, *premis Nobel d'economia*. Deixant de banda la discussió de si és millor mantenir-ho junt o separat, el que sembla prou evident és que si s'utilitzen els dos primers ja no cal el tercer.

3.4.4 Especificitat i exhaustivitat

El nivell d'especificitat que s'ha trobat és, en termes generals, adequat. Normalment, i amb ajuda dels termes compostos, s'assignen paraules clau que representen de manera bastant concreta i específica el contingut del document. Això es dóna més habitualment en les pàgines d'especialització temàtica, i no tant en les de contingut més genèric.

Tot i que s'aconsella ser tan específic com sigui possible en la indexació, no es pot oblidar que, igual que passa amb l'exhaustivitat, el nivell d'especificitat ha de ser adequat al nivell expressat pels usuaris en la recuperació. Si la indexació és específica però les cerques dels usuaris no ho són tant, tindrem un problema important de silenci, més encara en un sistema no controlat que, per tant, no utilitza referències per mostrar relacions jeràrquiques entre conceptes. Probablement per aquesta mancança del sistema s'han detectat casos en què s'assignen alhora termes genèrics i termes específics per tal d'assegurar que la pàgina es recuperi tant en una cerca específica com en una de més general. En molts casos però, no es tracta de casos de poca especificitat, atès que la paraula clau més genèrica assignada assoleix el nivell d'especificitat adequat per al contingut del document. El que es fa és assignar un descriptor que representa el concepte genèric del document i, al mateix temps, descriptors que corresponen a conceptes més específics que apareixen com a part de la pàgina. Per exemple, en una pàgina de recursos sobre *tractament de residus*, a part de *tractament de residus*, poden aparèixer descriptors com *aigües residuals*, *residus tòxics*, *residus orgànics*, *reciclatge*, etc.

Això ens porta a parlar d'exhaustivitat, ja que aquests termes específics corresponen a la matèria d'alguns dels recursos que apareixen llistats a la pàgina. És similar a la representació de contingut d'un document en què s'indexa el sumari complet, només que en aquest cas s'afegeix també a la representació el concepte més genèric que expressa conjuntament tot el contingut del document.

Es pot parlar de sumari d'un document en el cas de pàgines web constituïdes per enllaços a altres documents? En un document tradicional amb sumari, es pot indexar el sumari per entendre que reflecteix el contingut del mateix document. Els conceptes específics representats en el sumari es troben, en principi, tractats al mateix document. En realitat, aquí ens trobem amb una nova situació en la qual el document, la pàgina web, és únicament i exclusiva el sumari. Els enllaços que conté porten a documents diferents, a pàgines que no formen part del que s'està indexant i que no seran indexades en el mateix sistema si són externes. Això vol dir que la informació que contenen les pàgines de destinació proporcionades a partir dels enllaços es perd, i no es pot recuperar en les cerques fetes dins del sistema, si no es representen de manera exhaustiva i específica els conceptes de què tracten.

En pàgines que no tracten de matèries especialitzades relacionades amb disciplines acadèmiques, el grau d'exhaustivitat assolit és més alt i, en alguns casos, del tot excessiu. En una pàgina relacionada amb el préstec interbibliotecari, com pot ser un formulari de sol·licitud de document, s'han trobat descriptors com ara *tarifes*, quan en realitat hi ha una pàgina específica amb les tarifes o preus del servei, informació que no apareix en la pàgina que conté el formulari de sol·licitud, si no és en forma d'una frase que diu “Consulteu les tarifes vigents”.

En línies generals, es pot dir que el grau d'exhaustivitat és relativament alt en la majoria de pàgines i, en alguns casos, hi ha certa tendència a no quedar-se amb el concepte més específic que abasta tot el contingut del document en el seu conjunt, sinó que es representen també matèries més específiques de les quals la pàgina no tracta realment sinó que en dóna notícia o hi fa referència d'alguna manera.

Imaginem una pàgina on es llisten tots els serveis de la Biblioteca, però sense aportar informació específica de cap servei, només en forma d'enllaços que porten a pàgines amb informació de cada un dels serveis. Si es posa el nom de tots els serveis llistats a la pàgina en el camp de paraules clau, quan l'usuari faci una cerca per *fotocòpies gratuïtes* obtindrà, a més a més de la pàgina específica on es dóna tota la informació sobre el servei en concret, la pàgina on es llisten tots els serveis, document que en realitat no conté cap informació sobre el servei de fotocòpies gratuïtes. Com tots sabem, aquest resultat, en la recuperació, té un nom: soroll.

És justament el cas contrari al que s'ha comentat més amunt, referit a la indexació de la pàgina sobre *tractament de residus*. Si som estrictes i decidim que la pàgina tracta de *tractament de residus* i no tenim en compte *aigües residuals* en la indexació perquè només es refereix a un dels recursos llistats a la pàgina, quan un usuari busqui aquest últim concepte no obtindrà cap resultat. És cert que no hi ha dins del sistema cap pàgina que tracti d'aigües residuals, però la de tractament de residus hagués ofert una possibilitat d'anar a parar finalment a una pàgina sobre el tema concret que interessa. Podríem parlar aquí de silenci? Si més no, de pèrdua d'informació segur que sí. Sobretot, si tenim en compte que la intenció de la pàgina és donar notícia i oferir accés als recursos concrets que hi apareixen.

Aquesta contradicció fa pensar en la possibilitat de tenir en compte un tractament diferent de l'exhaustivitat i l'especificitat en funció del tipus de pàgina que es vol indexar.

3.4.5 Coherència en la indexació

Resulta difícil trobar un mínim de coherència entre indexadors i, fins i tot, entre feines diferents d'un mateix indexador, quan no es fa servir un llenguatge documental comú i no es fa ni un mínim control de vocabulari. Més encara quan els documents que cal indexar no tracten d'un tema específic al voltant del qual hi ha una terminologia establerta i coneguda, sinó que tracten de temes diversos no sempre inclosos en disciplines acadèmiques.

En l'assignació de metadades de la BUB feta per responsables de les diferents disciplines científiques s'observa una diferència important referida a totes les qüestions comentades fins aquí. Aquestes diferències no es poden atribuir a la disparitat de recursos per indexar que es pot donar entre unes disciplines i altres, ja que l'estructura de pàgines web i la tipologia de recursos que s'hi inclouen és comuna i molt similar en tots els casos, especialment des que es va implementar el nou web.

Per tant, es pot suposar que l'establiment de pautes i criteris comuns, sumada a l'elaboració d'eines de control bàsic de vocabulari, hauria de contribuir a disminuir les diferències i, per tant, augmentar la coherència de la indexació.

3.5 Anàlisi de la recuperació

3.5.1 Abast i limitacions

L'anàlisi transaccional pot oferir una gran quantitat d'informació a partir de dades explotades per programes adequats de manera quantitativa. La interpretació qualitativa d'aquestes dades afegeix encara més informació que la que ja donen els números per si sols en molts casos.

Diferents autors d'estudis similars abasten altres qüestions que cal analitzar i en les quals cal entretenir-se a l'hora d'exposar les característiques de la recuperació. No és habitual trobar recollits en un únic estudi tots els elements informatius possibles que es poden arribar a extreure d'una anàlisi d'aquest tipus. Frías i Martín (Frías, 1999, p. 429–432) ofereixen una llista prou extensa de la tipologia d'informació que es pot extreure de l'anàlisi transaccional. En el treball de Jansen, Spink i Saracevic (Jansen, 2000, p. 209–211), es divideixen els elements informatius en tres grups: informació sobre les sessions, informació sobre la forma de les expressions de cerca i, finalment, informació sobre els termes utilitzats en la cerca. Cadascú, doncs, utilitza els elements que necessita i els agrupa i interpreta segons l'objectiu que s'hagi proposat per a l'anàlisi que duu a terme.

Hi ha moltes dades que nosaltres no hem examinat en aquesta avaluació, no per manca d'interès, que en tenen molt, sinó per limitacions de temps i d'eines i perquè no es corresponien amb els objectius inicials de l'avaluació. Algunes de tan interessants com el que Frías i Martín anomenen “frustració de l'usuari”, que es pot deduir a partir de l'anàlisi de la quantitat d'intents que fa un mateix usuari davant de resultats negatius —“repetició incrèdula”— i la manifestació de la pròpia frustració en forma de “grafiti en línia” o paraules malsonants i frases malicioses. El treball de Jansen també inclou aquest tipus de dades, la quantitat de cerques idèntiques, la quantitat i forma de les modificacions a partir d'una cerca inicial, etc.

En el cas de l'estudi de les cerques efectuades al web de la Biblioteca, l'objectiu de l'anàlisi és establir la relació amb els resultats de l'anàlisi de la indexació per tal de corregir-la, pautar-la i adaptar-la al màxim a les necessitats de recuperació dels usuaris.

A part de l'objectiu que guia l'avaluació, com ja s'ha esmentat més amunt, no es disposa de cap programa d'explotació de fitxers Log, i aquest fet també determina quins són els elements que es podran analitzar i amb quines limitacions. Tanmateix, el tractament de les dades que s'ha fet amb el programa *Excel*, sumat a un examen detallat de mostres extretes del conjunt total de dades, ha permès fer-se una idea bastant exacta del que busca l'usuari en el web de la Biblioteca, de com ho fa i de quins són els resultats que obté.

3.5.2 Dades generals

Taula 2. Dades quantitatives generals corresponents a les transaccions de maig i juny de 2002[10]

	Maig	**Juny**	**Total**	
Nombre de cerques	12.012	9.279		21.291
Adreça IP (*Internet Protocol*) de la UB	4.243	4.235	40%	8.478
Errors de sistema	449	176	3%	625
Cerques a metadades	11.427	8.874	95%	20.301
Cerques al text complet	136	229	2%	365
Accés a opcions de cerca	1.827	1.297		3.124

La variació en el nombre de cerques dels dos mesos confirma el que apuntàvem del menor ús del cercador en època d'exàmens. Malgrat que seria necessària una comprovació més extensa d'aquest punt per recolzar l'afirmació, la diferència d'utilització de les biblioteques universitàries en època d'exàmens respecte d'altres períodes és un aspecte prou conegut per tot el personal d'aquests centres. És factible, doncs, pensar que en el cas de la utilització del web també hi ha un efecte visible de la diferència.

El petit percentatge d'errors del sistema, molt més elevat al mes de maig que al de juny, és degut a causes tècniques o a característiques de funcionament del cercador que no és necessari explicar aquí, però sí que ens interessa conèixer aquesta dada i tenir-la present a l'hora de comptabilitzar altres aspectes, ja que en la majoria de casos caldrà restar-la del total de cerques que cal examinar.

En un 40% dels casos, les cerques s'han fet des d'ordinadors ubicats a la mateixa Universitat. El fet que la quantitat de cerques externes sigui encara més alta no es pot prendre com a indicatiu d'usuaris externs a la comunitat; en realitat resulta impossible saber quins d'aquests accessos externs corresponen a usuaris de la Universitat i quins no, ja que no cal identificar-se per accedir al web de la Biblioteca i fer cerques. Resulta curiós, tanmateix, que el nombre de cerques fetes des de la mateixa UB durant els dos mesos sigui tan similar quan el nombre de cerques totals mostra una clara variació d'un mes a l'altre.

L'elevada proporció de cerques fetes exclusivament al camp de metadades, el 95%, i la baixa quantitat de cerques fetes al text complet són degudes probablement a un motiu més relacionat amb les característiques de la interfície que amb la voluntat real de l'usuari. La cerca per defecte es fa als camps de metadades, no es pot saber que hi ha l'opció de fer la cerca al text complet dels documents si no es va expressament a la pàgina que conté les opcions de cerca. Per aquest motiu, ens ha semblat interessant comparar el nombre de visites que ha tingut aquesta pàgina d'opcions de cerca amb el nombre de vegades que s'ha triat el text complet. El resultat d'aquesta comparació indica que un 12% de les visites a la pàgina d'opcions de cerca es van aprofitar per modificar l'opció i buscar en el text complet.[11]

3.5.3 Com es fan les cerques

Taula 3. Dades referides a la manera en què es fan les cerques[12]

	Maig	Juny	Total	
Cerques per un únic terme simple	5.660	4.492	10.152	49%
Cerques per més d'un terme simple	5.903	4.611	10.514	51%
Cerques per terme compost/frase[13]	87,2%	84,8%		86%
Cerques de noms propis[14]	24,4%	30,3%		27,4%
Cerques de títols	13,6%	11,3%		12,5%
Operació AND	12,8%	15,2%		14%
Operació OR	154	133	287	1,4%
Operació NOT	3	4	7	0,3%
Operador frase ("")	79	80	159	0,8%
Operador truncament (*)	23	21	44	0,2%
Accessos a pàgina d'ajuda	143	115		258

Les cerques amb un únic terme simple representen gairebé la meitat del total de cerques vàlides. Aquesta xifra és bastant més alta que l'obtinguda a l'estudi de Jansen i els seus col·legues (Jansen, 2000, p. 214–216), que se situa en el 31% i que l'autor declara que és similar a l'obtinguda en altres estudis.

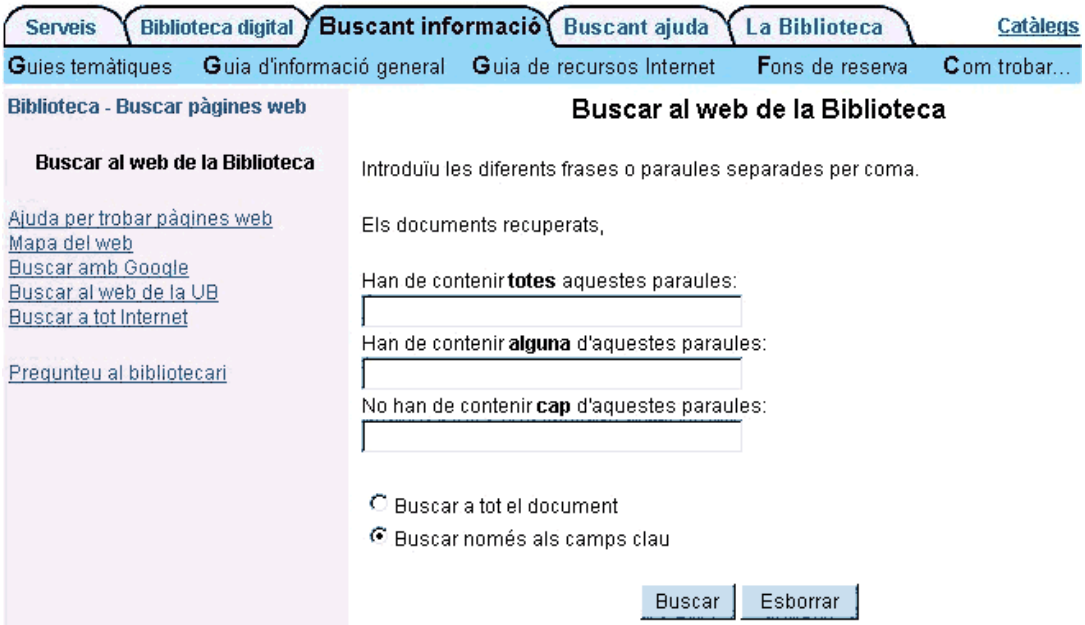
Pel que fa a les cerques amb més d'un terme simple poden ser de dos tipus:

a. termes simples o compostos combinats amb operadors booleans,
b. termes compostos, expressions o frases complexes entrades de manera directa.

Abans de comentar les dades sobre la utilització d'operadors, és convenient d'observar les pantalles de cerca i fer algun comentari respecte del seu funcionament.

Imatge 1. Casella de cerca present a la capçalera de totes les pàgines

A la capçalera de totes les pàgines web de la Biblioteca apareix una casella de cerca amb un enllaç a sota per consultar les opcions de cerca. La majoria d'usuaris introdueixen els termes directament sense anar a veure les opcions disponibles. L'operador booleà que es pot utilitzar en aquesta casella és el d'intersecció —AND—, però no hi ha cap indicació que, per fer això, els diversos termes s'hagin de separar amb comes. Si s'introdueixen diversos termes separats per un espai en blanc la cerca es fa per expressió o frase (cerca per cadena de caràcters).



Imatge 2. Pàgina d'opcions de cerca que permet utilitzar els operadors booleans

Tal com es veu en la segona imatge, per fer una cerca que contingui un terme *i* un altre (operador AND) cal posar-los junts en la primera casella, i per fer una cerca que contingui un terme *o* un altre (operador OR), cal posar-los junts en la segona casella. En aquesta pàgina d'opcions sí que hi ha una clara indicació de la necessitat de separar mitjançant comes els termes que es volen combinar.

Si comptabilitzem les vegades que s'ha omplert la segona casella, podem dir que l'operador OR s'ha utilitzat en 287 cerques. Ara bé, hem examinat amb detall totes aquestes cerques i hem trobat que en 194 casos només s'inclou un terme de cerca (simple o compost), és a dir, no hi ha en la mateixa casella cap altre terme amb el qual fer l'operació OR. Entre aquests 194 casos amb un únic terme de cerca, n'hi ha 70 en els quals s'ha escrit un altre terme en la casella de l'operador AND. Suposem que erròniament els usuaris es pensen que les diferents caselles són per introduir diferents termes de combinació, sense que tinguin massa clar amb quin tipus d'operador. La majoria de les 70 cerques en les quals s'ha trobat un terme a cada casella, tenen intenció d'intersecció, és a dir, els termes haurien d'anar junts en la primera casella separats per comes. En menys casos la intenció era utilitzar l'operador OR, per la qual cosa els dos termes haurien de ser en la segona casella.

Uns exemples de cerques d'aquest tipus poden ajudar a veure l'error que acabem de comentar:

- *Perfumeria* (casella AND) i *cosmètica* (casella OR): en aquest cas, per fer-ho correctament, s'haurien d'haver posat els dos termes separats per comes a la casella OR.
- *Història* (casella AND) i *medieval* (casella OR): aquí els dos termes haurien de ser a la casella AND, o bé separats per coma si es volien combinar; o simplement per un espai en blanc, si es volia trobar la forma composta *història medieval*.
- *Sistema fiscal* (casella AND) i *Noruega* (casella OR): sembla clar que hauria de ser una intersecció i, per tant, és igual que el cas anterior.

Respecte dels casos en els quals s'ha fet servir més d'un terme en la segona casella, hi ha una intenció clara i conscient de combinar-los amb l'operador OR, la quantitat obtinguda és de 47 (sense comptar les cerques idèntiques). Només en 15 d'aquestes expressions de cerca els termes estaven correctament separats per comes. En la resta de casos estaven separats per espais en blanc.

En resum, es pot dir que de les 287 vegades que s'ha utilitzat l'operació OR (1,4% del total de cerques), només en 47 casos (16%) hi ha hagut coneixement real del que s'estava fent, i només en 15 (5%) s'ha arribat a fer de manera correcta. El percentatge d'error en la utilització d'aquest operador es pot situar, doncs, en el 95%.

Pel que fa a l'operador NOT, 7 casos de 20.666 cerques no semblen suficients per fer cap altre comentari que no sigui l'aparent invisibilitat per a l'usuari. Tanmateix, afegirem que en 4 d'aquests casos les caselles AND, OR i NOT contenien exactament el mateix terme. En els altres 3, hi

havia un terme diferent a cadascuna de les tres caselles. Tot i que resulta difícil de saber amb seguretat, exemples com ara *tesi* AND *iber* OR *vestits* NOT, tenen tota l'aparença de voler ser en realitat interseccions amb l'operador AND. En els tres casos es té la mateixa impressió. Aquestes xifres podrien situar el percentatge d'error en la utilització d'aquest operador en el 100%.

Quant a la utilització de l'operador AND, no resulta fàcil d'analitzar. La dificultat rau principalment en el fet que la voluntat de combinació no s'expressa correctament, ja que habitualment la cerca es fa des de la capçalera de qualsevol pàgina, sense passar per la pàgina d'opcions on hi ha l'explicació sobre la separació de termes mitjançant comes. Es fan servir indistintament 'i', 'y', el signe '+', el símbol '&', el terme 'AND', o, en la majoria de casos, simplement s'escriuen diversos termes separats per un espai en blanc. Creiem que els usuaris copien els diferents sistemes dels principals cercadors genèrics existents a la web. Si un usuari està acostumat a utilitzar *Altavista*, prova de fer la combinació escrivint AND o amb el modificador + davant del terme; si està acostumat a *Google*, ho fa només amb un espai en blanc, etc. Tampoc no es fan servir en gaires ocasions les cometes per indicar la cerca de frases o expressions senceres. Amb tot, això no és greu pel que fa als resultats, atès que la cerca per defecte quan hi ha termes separats per un espai en blanc és de cadena de caràcters i, per tant, el resultat és el mateix que si es posen les cometes. Llavors, l'única manera de distingir termes compostos, expressions o frases, de combinacions de termes amb operació AND, és l'examen detallat de cada expressió de cerca. A diferència del que passava amb els operadors OR i NOT, l'elevat volum de cerques que caldria examinar en aquest cas fa que la tasca es presenti d'entrada com a inabastable. La solució adoptada ha estat triar una mostra d'un 5% de les cerques de cada mes amb més d'un terme per fer-ne l'anàlisi detallada i poder oferir un tant per cent aproximat.

En general es pot dir que la majoria de cerques en què hi ha més d'un terme corresponen a frases o a termes compostos (86%); en pocs casos (14%) es poden associar a l'ús de l'operador AND per combinar conceptes. Entre els casos en què s'aprecia una intenció clara de fer la combinació, només n'hem trobat 1 amb els termes separats per comes tal com requereix el programa. En la resta, els diversos termes (simples o compostos) estan separats per un espai en blanc, excepte en tres casos en què s'ha escrit explícitament l'expressió AND i dos en què s'ha utilitzat el signe + davant dels termes.

L'usuari busca de forma precoordinada, seguint el costum del llenguatge natural. És normal trobar cerques com ara "revistas y periódicos que hablen de Carlos V", "información de todo lo que es las funciones de las bibliotecas", o "revistas electrónicas especializadas en tecnología de la información", tot i que també hi ha molts casos de termes compostos formats per substantiu + adjectiu.

Una dada que crida l'atenció és la important quantitat de noms propis (autors, personatges diversos i personal de la institució) que es busquen. De les cerques formades per termes compostos i frases, un 24% corresponien a noms propis. D'altra banda, el percentatge de títols de documents també és prou important (17%). L'usuari no és conscient de la diferència que hi pot haver entre buscar pàgines web i buscar documents en un catàleg o en una base de dades. De cerques tan curioses com ara *porno maduros*, *seguros de viaje*, *listas de música dance*, *conservas de espárragos*, etc., es pot deduir que l'usuari no sembla que tingui gaire idea de què pot trobar en el lloc web d'una biblioteca universitària.

Tot i l'abundància de cerques per frases o expressions complexes, poques vegades es fan servir cometes per indicar aquest tipus de cerca, tot i que ja hem dit que a efectes pràctics no és important. La quantitat d'intents de truncament també és mínima. Altres capacitats del sistema explicades a la pàgina d'ajuda, com ara l'operador CASE per a cerques exactes sensibles a majúscules i accents, ja ni apareixen registrades en els fitxers Log.

Aquestes dades tan baixes d'utilització d'operadors i símbols per limitar, concretar o ampliar la recuperació, però tan altes quant a errors i desconeixement del sistema, coincideixen amb les trobades per Jansen et al. i també, segons els mateixos autors, amb les obtingudes per altres treballs. Aquests autors destaquen, respecte d'aquestes dades, la diferència amb els resultats que es troben habitualment en estudis de la recuperació efectuats en sistemes especialitzats.

Una dada que pot ser interessant d'observar, vista la taxa d'errors i el gran desconeixement de les capacitats i mecanismes de cerca del sistema, és la quantitat d'accessos a la pàgina d'ajuda en la cerca, disponible des de la pàgina d'opcions de cerca. A la pàgina esmentada <http://www.bib.ub.es/bub/ajuda_buscar.htm> s'explica el funcionament dels operadors, la diferència entre la cerca a les metadades i al text complet, i altres funcionalitats del cercador com ara l'ús de cometes i de l'operador CASE. La quantitat de visites a la pàgina d'ajuda entre els dos mesos és només de 258. Aquesta dada és prou baixa per recolzar les afirmacions expressades tantes vegades per responsables de llocs web: els usuaris no es llegeixen els textos explicatius ni les ajudes; tal com exposen també altres autors (Jansen, 2000, p. 217), la major part de persones actuen segons el mètode conegut com a *assaig i error*. No s'ha comptabilitzat de manera rigorosa, però l'examen d'altres aspectes ha permès observar que hi ha un nombre elevat de casos de cerques fetes en castellà i una quantitat molt petita de cerques en anglès (exceptuant les cerques corresponents a títols de revistes científiques). Això corrobora la

necessitat d'establir les equivalències interlingüístiques entre descriptors, però dóna prioritat a la forma castellana respecte de l'anglesa.

3.5.4 Quins resultats s'obtenen

És interessant saber com busquen els usuaris i és indubtable que l'anàlisi de la seva actuació aporta molta informació valuosa, però potser encara és més útil conèixer els resultats de les cerques que duen a terme. El bon o mal resultat és el que en realitat pot donar una mesura de l'adequació de la indexació que es fa.

Taula 4. Dades referides a resultats de cerca

	Maig	**Juny**	**Total**	
Cerques amb resultat = 0	9.860	7.578	17.438	84,4%
Cerques amb resultat 1–5	1.011	875	1.886	9,1%
Cerques amb resultat 6–10	177	153	330	1,6%
Cerques amb resultat > 10	515	497	1.012	4,9%

Donada la naturalesa de les pàgines web de la Biblioteca i coneixent-ne el contingut, els resultats "ideals" serien els que se situen entre 1 i 5 documents. Aquesta idoneïtat, però, implica, a més d'una bona indexació, que l'usuari fa les cerques correctament. Si es donessin les dues condicions no haurien d'aparèixer més de 5 documents com a resultat de cerca, màxim 10, atès que no és normal que hi hagi més quantitat de pàgines referides a un mateix tema. Hem vist que aquests resultats, teòricament ideals, només es donen en un 9,1% dels casos.

El primer que crida l'atenció en veure la taula 4 és l'elevat volum de cerques amb resultat 0. La primera temptació, sabent que la majoria de cerques es fan exclusivament al camp de metadades, és modificar la configuració del cercador per tal que faci la cerca al text complet de les pàgines. D'aquesta manera es reduiria de manera important la quantitat de resultats negatius. El soroll que això pot provocar és important i, probablement, encara hi hauria molts resultats 0. Sabem que és molt frustrant per a l'usuari no trobar res, però no creiem que oferir-li informació no rellevant sigui adequat per disminuir la seva frustració. Així doncs, cal analitzar els motius dels resultats 0, saber si corresponen a casos de silenci documental i provar d'esmenar el que calgui per tal de reduir-lo. Només hi ha un camí per conèixer els possibles motius d'aquests resultats: l'examen detallat de les cerques. Davant de l'elevat volum de dades per examinar cal treballar amb mostres. Concretament, s'ha analitzat una mostra equivalent al 5% dels casos que donen 0 documents de cada mes.

A partir de l'estudi de les mostres es constata que hi ha diferents motius que provoquen els resultats de les cerques que hem obtingut. Alguns estan relacionats amb la forma dels termes i la indexació, d'altres amb la construcció de les expressions de cerca i d'altres, la majoria, són deguts purament i simplement a la crua realitat: no hi ha documents sobre el tema buscat.

Més concretament, els tres motius més freqüents de l'obtenció de 0 documents són:

- Es busquen temes que no existeixen al web de la Biblioteca: "Conveni col·lectiu de l'Hotel Melià Barcelona S.A.", "en qué momento y por qué se creó el día internacional del niño", "demostrar que la media aritmética es mayor que la media geométrica", "Mortadelo y Filemón", etc. Són exemples de cerques que, sigui quina sigui la indexació, seguirien donant com a resultat 0 documents. Això no es pot considerar un silenci documental, ja que no es tracta de documents rellevants que no apareguin als resultats de cerca, sinó que la realitat és que la Biblioteca no té cap informació sobre aquests temes. Ho trobem en un 27% de les cerques examinades en la mostra. Es busquen noms propis o títols de documents que no apareixen específicament a les pàgines web. Aquesta casuística es dóna en un 31% de la mostra examinada. Es dóna la circumstància que molts d'aquests noms i títols es trobarien si es busquessin al catàleg o, en els casos de personal docent i administratiu de la institució, al directori X500 del mateix web. És a dir, en realitat la Biblioteca disposa de la informació sol·licitada, però cal buscar-la en altres llocs.
- Es busquen termes massa específics en un 36% de les cerques (descomptant-hi noms i títols). No hi ha pàgines que tractin d'aquests temes tan específics, aquesta és la realitat. Malgrat tot, hi ha pàgines que els inclouen o abasten d'una manera o altra, o bé que ofereixen enllaços adequats a altres webs específics sobre aquests temes. Són casos en què les pàgines de la Biblioteca haurien pogut arribar a solucionar la necessitat d'informació si la cerca s'hagués enfocat d'una manera més genèrica o la indexació hagués estat més exhaustiva.

Per exemple, s'ha buscat *AACR* i també *CDU*. No hi ha cap pàgina específica de cap dels dos

temes, però sí que hi ha, a la guia temàtica de biblioteconomia, una pàginadedicada a normes, un recull d'eines per al catalogador i una pàgina amb tutorials, entre els quals n'hi ha un sobre la CDU.

De la mateixa manera, no hi ha cap pàgina dedicada a *Jean Piaget*, però, a la guia temàtica d'educació, se'n pot trobar una sobre educadors cèlebres <http://www.bib.ub.es/www5/5edu17.htm>, entre els quals hi ha Jean Piaget i s'ofereixen enllaços a obres d'ell, obres que tracten d'ell i webs especialitzats en ell. El mateix passa amb *educador Vigotsky*, un altre terme buscat sense èxit i present a la mateixa pàgina sobre educadors.

En aquests exemples que hem posat, hi ha dues solucions possibles que haguessin evitat el silenci: o bé es fa la cerca a text complet (recordem que ho ha de triar l'usuari anant a la pàgina d'opcions de cerca), o bé es fa una indexació exhaustiva per a les pàgines que llisten recursos, possibilitat que ja hem comentat abans.

En els casos de cerques de conceptes relacionats amb les ciències experimentals i de la salut, el grau d'especificitat és més alt encara i, ni la cerca a text complet, ni la màxima exhaustivitat en la indexació solucionarien el problema de manera totalment satisfactòria.

El 6% restant de resultats iguals a 0, és degut a diferents motius, entre els quals trobem errors tipogràfics o ortogràfics, mala formulació de l'expressió de cerca, assignació incompleta de metadades a les pàgines, cerques fetes amb els termes en castellà, i un petit nombre de cerques que no hem pogut entendre, ja que consten d'un conjunt de números i no sabem a què corresponen.

Un exemple interessant de silenci, relacionat amb un títol i un nom propi, però a causa d'un problema en l'expressió de cerca, és “El Cristo de Velázquez”. Es dóna la circumstància que hi ha, a la guia temàtica de recursos d'art, un conjunt de pàgines dedicades de manera monogràfica al pintor Diego Velázquez <http://www.bib.ub.es/velazquez/1vportada.htm>. En aquest recull monogràfic es poden trobar imatges de quadres de Velázquez (entre ells el del Crist), bibliografia, enllaços a bases de dades i catàlegs amb la cerca per Velázquez ja elaborada, enllaços a altres webs sobre el personatge i a articles de premsa. La pàgina que conté la imatge del Crist incorpora les metadades adequades: *Cristo Crucificado*, *Diego Velázquez*, etc., però l'usuari ho ha cercat incorrectament perquè no sabia el nom del quadre o no ha fet la combinació d'intersecció amb els termes *Cristo* i *Velázquez*.

Un exemple de silenci provocat per l'idioma és la cerca pel terme *mapas*. Si s'hagués buscat la forma catalana o hi hagués les equivalències de metadades en castellà, s'haguessin trobat 10 documents sobre mapes, el primer dels quals és una pàgina on s'explica com trobar mapes i atles al catàleg, a bases de dades o al web; els 9 següents són pàgines amb recursos sobre mapes de biologia, de geologia, de medi ambient, històrics, topogràfics, etc. Un altre exemple similar és la cerca pel terme *medio ambiente*, que ha donat 0 documents mentre que si es fa en català se n'obtenen 15.

El fet de no separar les interseccions amb comes ha fet que el cercador busqués termes compostos o expressions quan en realitat es volia fer una combinació AND. Si s'hagués fet correctament la combinació el resultat no hauria estat de 0 documents en una part dels casos.

Respecte de les cerques amb més de 10 resultats, sembla que el problema està provocat per la conjunció de diversos motius. Un d'aquests és el contingut de la frase resum de l'etiqueta “Description”. Hem de pensar que la cerca a les metadades no es fa només al camp de paraules clau, sinó també al camp de descripció. En aquest camp hi apareixen termes que es repeteixen a moltes de les pàgines formant part de la frase de descripció, tot i que no s'han considerat representatius del contingut i no consten com a descriptors. Per exemple, a moltes pàgines s'acaba la frase de descripció amb la cadena “...de la Biblioteca de la Universitat de Barcelona”. Això fa que, en buscar informació sobre la mateixa Biblioteca mitjançant el terme *biblioteca* o *biblioteca de la Universitat de Barcelona*, apareguin un munt de resultats que no corresponen a pàgines on es parla realment de la Biblioteca.

El mateix passa amb termes com ara *Internet*, *recursos*, etc., termes que apareixen a totes les pàgines on es llisten recursos disponibles al web, o termes referits a forma, com ara *revistes*, *diccionaris*, etc. Si la cerca es fa sense combinar amb cap altre terme, com hem vist que passa en el 49% de les cerques, el resultat és forçosament elevat. Així per exemple, si es busca *diccionaris*, o *bases de dades*, o *revistes*, els resultats són excessius perquè apareixen totes les pàgines que contenen diccionaris sobre cada tema, bases de dades de cada matèria i revistes sobre cada disciplina. Són rellevants, però són excessius. Tots aquests resultats apareixen barrejats amb les pàgines referides a la forma com a contingut, com ara la de “Com trobar revistes?” o “Preguntes més freqüents sobre bases de dades”, que queden totalment amagades entre pàgines i pàgines de resultats que l'usuari no s'arribarà a mirar perquè són excessives.

També es detecta un aspecte de mala pràctica en la indexació que fa aparèixer soroll en alguns casos. Es tracta de fer constar la matèria genèrica en pàgines de contingut més específic. Per

exemple, el descriptor *art* apareix en la majoria de pàgines que formen *IMAGO*, un arxiu d'imatges d'art <http://www.bib.ub.es/bub/imatges/artistes.htm>. En aquest arxiu es poden trobar imatges amb obres de moltíssims artistes. Si es fa una cerca per *art*, surten centenars de resultats que porten a pàgines dedicades a artistes de manera individual, quan en realitat, només obtenint la pàgina principal d'entrada a *IMAGO* n'hi hauria prou.

Passa el mateix quan es fa una cerca pel terme *medicina*: apareixen més de 50 resultats, entre els quals hi ha pàgines sobre odontologia i sobre infermeria, disciplines que tenen totes dues reculls de pàgines pròpies, malgrat que estan relacionades i comparteixen recursos amb la guia temàtica de medicina.

Hi ha també un nombre (no excessivament important) de cerques fetes per articles, preposicions, etc. Una cerca feta amb *de*, per exemple, ofereix més de 2000 resultats, i també passa el mateix fent-ho amb *a*, *d*, etc. En aquest sentit, s'hauria de modificar la configuració del cercador perquè no busqués termes amb menys d'un cert nombre de caràcters i afegir un fitxer de paraules buides.

La conjunció de cerques per termes únics fetes per l'usuari, una indexació que utilitza matèries genèriques i específiques alhora per assegurar resultats, i l'aparició de termes genèrics al camp de descripció, són els factors principals que provoquen el soroll que hem observat. Tanmateix, no pensem que aquest sigui un problema excessivament greu quant a volum, al contrari del que passa amb els resultats 0. Els documents resultants són rellevants, encara que n'hi hagi més dels necessaris i la inclusió del contingut del camp “Description” a sota de cada document resultant ajuda a destriar ràpidament el que interessa del que no.

Malgrat l'elevat grau de rellevància que hem observat, ja que els documents resultants tracten del tema expressat en la cerca, encara caldria fer-se una altra pregunta: són pertinents o realment útils aquests resultats per a l'usuari? La resposta no la sabem. Tal com ens recorden Frías i Martín (Frías, 1999, p. 432), aquest és un dels inconvenients de l'anàlisi transaccional: no podem saber en realitat què necessitava l'usuari, ni quins resultats triarà com a més útils; només sabem com ha expressat aquesta necessitat i què li hem ofert per satisfer-la.

4 Conclusions

Un cop analitzada la indexació de les pàgines web amb metadades i el resultat d'utilitzar-les en la recuperació, arriba l'última part de l'avaluació de les metadades del web de la Biblioteca: a partir del que s'ha vist, cal fer una sèrie de propostes que poden ajudar a millorar el sistema de metadades com a eina d'indexació i recuperació del web.

- **Format i quantitat de metadades**: pensem que el millor és continuar utilitzant un format genèric en lloc d'un específic com ara Dublin Core. Els cercadors generals no el reconeixen i, de fet, el resultat és el mateix. En cas de necessitat en un futur, sempre es pot configurar en el cercador una equivalència que li faci entendre d'igual manera “DC.Description” que “Description”. No es considera tampoc útil afegir més elements o camps de metadades, ja que les necessitats de recuperació i el tipus de documents inclosos al web de la Biblioteca no ho fan necessari.

- **Singulars i plurals**: creiem que concretant i posant per escrit les recomanacions de la norma UNE 50-106-90 n'hi ha prou. El tema ha funcionat bastant bé sense marcar pautes prèvies, per pura intuïció que prové del coneixement i ús del llenguatge natural. Pel que fa als usuaris en la fase de recuperació també, en línies generals, han buscat en singular conceptes que es representaven en singular i el mateix en el cas del plural.

- **Categories gramaticals**: és necessari substituir els adjectius i verbs per formes substantivades i establir clarament les pautes perquè es faci sempre així. Pel que fa a l'anàlisi de la recuperació, en poques ocasions els usuaris han utilitzat expressions de cerca consistents en adjectius, adverbis o formes verbals.

- **Sinònims i formes variants**: aquest aspecte necessita actuació immediata. Una proposta és avançar el treball amb els responsables del cercador per tal d'incloure-hi un diccionari de termes amb les equivalències corresponents. El personal indexador ha de poder consultar la llista de descriptors amb equivalències incloses, cosa que ajudarà a augmentar la coherència i estalviarà la feina de pensar en tots els possibles descriptors equivalents i escriure'ls. A més a més, el control de la sinonímia feta pel programa cercador mantindrà el conjunt actualitzat, sense haver de fer modificacions directament a les metadades de les pàgines.

  També cal normalitzar la forma dels noms propis. Atès l'elevat nombre de noms propis que es busquen, és important treballar aquest aspecte al qual en principi no s'havia donat una gran importància.

Finalment, en vista de la gran quantitat de cerques efectuades en castellà, és bastant urgent afegir les equivalències dels descriptors en aquest idioma.

- **Termes simples i termes compostos**: tot i que gairebé la meitat de cerques es fan amb un únic terme simple, creiem que és recomanable la utilització de termes compostos en la indexació en els casos en què el terme simple no es correspon a una matèria amb entitat pròpia. Hem vist que en la recuperació no es fan gaires combinacions de termes; quan se'n busca més d'un, es fa amb termes compostos o frases. Per tant, no té sentit mantenir termes simples com ara *codis*, *anys*, *volums*, *números*, *organismes*, etc.; pensant que ja els combinaran. L'usuari ni combina ni sap combinar: o busca un terme simple o bé una frase o un terme compost. Tenint en compte el sistema de cerca del programa, si busca un terme simple i aquest forma part d'una paraula clau composta ho trobarà de totes totes. És a dir, si busca "organismes" trobarà les pàgines sobre "organismes públics", "organismes internacionals", "organismes professionals"... Sempre que es pugui completar el sentit d'un concepte mitjançant un terme compost s'hauria de fer.

  És necessari també mantenir els termes simples que indiquen forma en el conjunt de descriptors, ja que l'usuari busca forma en la mateixa mesura que contingut. Caldrà prendre les mesures de formació necessàries perquè l'usuari concreti més les seves cerques, combinant aquest tipus de termes amb els que representen contingut.

- **Exhaustivitat i especificitat**: s'han detectat diferents nivells d'exhaustivitat, i no s'ha observat gaire coherència en aquest aspecte. La proposta que fem passaria per establir criteris diferents segons el tipus de pàgina que s'ha d'indexar. No es pot ser excessivament exhaustiu en pàgines amb informació genèrica, però cal augmentar el grau de detall en les pàgines especialitzades temàticament i, sobretot, en les que consisteixen bàsicament en llistes de recursos. Són els casos en què la Biblioteca no ofereix informació específica sobre una matèria però hi dóna accés a partir d'una pàgina de temàtica una mica més àmplia. En aquests exemples, potser es podria mantenir la indexació amb l'especificitat adequada per al conjunt de la pàgina en el camp de paraules clau i incloure els temes referenciats de manera més específica a la pàgina en el camp de descripció.

  També seria convenient evitar la doble assignació de termes específics i genèrics alhora (més genèrics que el contingut de la pàgina). Si una pàgina conté imatges d'obres de l'arquitecte Aalto, la indexació adequada des del punt de vista de l'especificitat ha de consistir en el nom del personatge, no s'haurien d'assignar també descriptors com ara *art* o *arquitectura*. Amb aquesta mesura disminuirien els resultats tan voluminosos que hem trobat en algunes cerques.

  Un aspecte que pot ajudar a evitar el problema del silenci relacionat amb cerques massa específiques amb el cercador, és evitar-les. Amb això volem dir que cal potenciar la navegació per l'estructura temàtica jeràrquica del web, i l'ús d'altres eines útils existents a la Biblioteca a l'hora de recuperar informació.

  Les guies temàtiques ofereixen informació especialitzada molt estructurada i aquesta informació està formada per una àmplia i diversa tipologia de recursos, des de pàgines web fins a registres bibliogràfics existents al catàleg o a bases de dades. Això proporciona una aproximació més genèrica a la informació, que permet a l'usuari arribar a nivells més específics de manera esglaonada. Els usuaris que no han trobat res quan han utilitzat el cercador per buscar coses tan específiques i concretes com ara *cortes geológicos* i *sistema renal*, tot i que haurien arribat a recuperar alguns recursos interessants sobre aquests temes si en lloc de fer una cerca haguessin navegat a partir de les guies temàtiques de geologia i medicina respectivament, des d'on se'ls hagués enviat també al catàleg i a les bases de dades adequades.

  El cercador de pàgines web és un sistema complementari de recuperació d'informació, que serveix per al que serveix, i cal optimitzar-lo, però creiem que la navegació per estructures lògiques i la utilització d'altres eines de recuperació més adequades a les diferents necessitats poden ser les millors solucions.

Altres qüestions que cal tenir en compte:

- **Modificació de la interfície de cerca**: tot i no estar relacionat directament amb la indexació, potser és un dels aspectes més importants que s'han de treballar per tal de reduir el fracàs en la recuperació. No es pot pretendre que al web hi hagi coses que no hi han de ser, però cal aconseguir la manera que l'usuari sàpiga quin tipus d'informació pot trobar amb la cerca de les pàgines web i quines altres eines té per buscar una tipologia d'informació diferent (directoris i catàleg per noms i títols). Potser es poden potenciar aquestes altres eines de recuperació d'informació des de la mateixa interfície que la cerca de pàgines de web. Així mateix, és important trobar la manera de simplificar el sistema de cerca i que el funcionament del cercador quedi clar sense necessitat de grans explicacions. El sistema de les tres caselles, una per a cada operació booleana i el fet que calgui accedir a la pàgina d'opcions de cerca per saber com funciona contribueixen a la

mala utilització i els consegüents mals resultats.

- **Modificació de la configuració del cercador**: pensem que la cerca per defecte no hauria de ser de cadenes de caràcters, sinó d'operació d'intersecció com és en la majoria dels grans cercadors. Quan l'usuari utilitza diversos termes separats per un espai, és molt difícil que els termes i l'ordre en què els escriu coincideixi exactament amb una expressió o frase localitzada als camps de metadades; mentre que, si es fes la combinació, hi hauria més possibilitats de trobar els diversos termes en una mateixa pàgina, repartits entre els camps "Description" i "Keywords". Sabem que es pot produir certa quantitat de soroll a causa de la falsa coordinació, però segurament no deu ser en una quantitat tan important com la de resultats 0 que es poden evitar. A més a més, amb aquesta mesura, se solucionarien els casos en què realment es vol fer l'operació AND i no es fa correctament, amb el consegüent silenci documental. Les cerques *estadístiques préstec i horaris biblioteca*, amb resultat 0, són exemples clars del que acabem de dir.

- **Formació** en les característiques del cercador a bibliotecaris i a usuaris: és important fer una campanya de formació per tal que els bibliotecaris, indexadors o no, coneguin a fons les característiques del cercador. Pel que fa als usuaris, potser es podria elaborar un tutorial senzill sobre com recuperar informació, ús d'operadors booleans, navegació per l'estructura temàtica del web, etc. Al mateix temps es podrien incorporar alguns exemples de diverses cerques en la mateixa pàgina de cerca.

Bibliografia

Asociación Española de Normalización y Certificación (1999). "Directrices para el establecimiento y desarrollo de tesauros monolingües, UNE 50 106:1990". En: Asociación Española de Normalización y Certificación. *Documentación*. 3ª ed. Madrid: AENOR, p. 77–126.

Frías, José Antonio; Martín Rodríguez, Fernando (1999). "El análisis transaccional como técnica de recogida de datos para el estudio del comportamiento de los usuarios en el catálogo en línea". En: Congreso ISKO-España EOCONSID'99 (4rt: 1999: Granada). *La representación y la organización del conocimiento en sus distintas perspectivas: su influencia en la recuperación de información*. Granada: Capítulo Español de la Sociedad Internacional para la Organización del Conocimiento : Facultad de Biblioteconomía y Documentación, p. 427–433.

Jansen, Bernard J.; Spink, Amanda; Saracevic, Tefko (2000). "Real life, real users and real needs: a study and analysis of user queries on the web". *Information processing and management*, no. 36 (2000), p. 207–227.

Notes finals

[1] Fins el passat 2 de setembre, responsable del web de la Biblioteca de la Universitat de Barcelona.

[2] Una mostra de treball sobre la utilització de metadades en un web amb informació genèrica en lloc d'un conjunt de recursos especialitzats es pot veure a: Assumpció Estivill (coord.), "Recursos web i metadades: informe de recerca", *BiD: textos universitaris de biblioteconomia i documentació*, núm. 7 (desembre 2001) <http://www.ub.es/biblio/bid/07estiv1.htm> [Consulta: 24 octubre 2002].

[3] Un exemple interessant d'estudi de la recuperació en entorns web genèrics és l'article de Jansen, Spink i Saracevic (Jansen, 2000). En aquest article es fa una anàlisi de les cerques efectuades al cercador Altavista a partir de l'estudi dels registres Log de les cerques.

[4] Es pot trobar un quadre comparatiu dels principals cercadors el maig del 1999, en el qual una de les qüestions que es fa constar és la capacitat de fer la cerca en els camps de metadades a: Santi Muxach, Ana Lopo, "Metadades a peu pla", *Item: revista de biblioteconomia i documentació*, núm. 24 (gener-juny 1999), p. 128-130.

[5] El programa era Search.cgi, de lliure distribució i descarregat de la xarxa.

[6] Passat un temps es va normalitzar la forma del nom de les seccions de la BUB per tal d'omplir el camp "Author".

[7] Tot es registra en un ordinador del Centre d'Informàtica de la UB, on s'ha instal·lat el cercador, i durant molt temps no hi ha ningú que se'n faci càrrec i a qui poder demanar aquesta informació.

[8] El nombre de pàgines que formen el lloc web no és fix, el més normal és que augmenti progressivament i que, de tant en tant, es faci neteja fusionant informació, eliminant informació obsoleta, etc.

[9] Per determinar aquest percentatge, s'ha examinat una mostra formada pels 100 primers termes simples del conjunt de metadades.

[10] Les proporcions s'han calculat respecte del total de cerques.

[11] La cerca al text complet inclou també els camps de metadades.

[12] Les proporcions s'han calculat després d'haver descomptat els errors de sistema, ja que les cerques amb error no han estat examinades.

[13] La impossibilitat d'examinar individualment totes les expressions de cerca amb més d'un terme ens ha dut a treballar amb una mostra del 5 % d'aquestes expressions. Per aquest motiu donem únicament el tant per cent i no quantitats totals i exactes en els casos de cerques per terme compost o frase i cerques amb operador AND. La resta de càlculs sobre operadors (OR, NOT, "", *) es donen respecte del total de cerques vàlides.

[14] Les proporcions de noms propis i títols es calculen respecte del total de cerques per termes compostos, expressions o frases, però no inclouen les operacions d'intersecció.

Facultat de Biblioteconomia i Documentació
Universitat de Barcelona
Barcelona, desembre de 2002
http://www.ub.edu/biblio • Comentaris

Citació recomanada • Metadades
UB • Facultat • BiD