Xarxa Punt TIC

MÒDUL 1 NIVELL AVANÇAT
Recerca i recuperació de la informació a Internet (avançat)
Apunts complets

plan avanza2 | GOBIERNO DE ESPAÑA | Plan E | MINISTERIO DE INDUSTRIA, TURISMO Y COMERCIO | SECRETARÍA DE ESTADO DE TELECOMUNICACIONES Y PARA LA SOCIEDAD DE LA INFORMACIÓN | Generalitat de Catalunya

➔ ÍNDEX

➔ A. Problemàtica documental de la informació al Web

Amb l'aparició dels BBS (sigla de *Bulletin Board System*), que permetia accedir de manera rudimentària a dades remotes usant un mòdem, l'accés a documents des d'una computadora casolana es va simplificar bastant. Les boques de producció de documents electrònics es van multiplicar, així com la quantitat d'usuaris amb computadora que intercanviaven informació. Tot i així, cada BBS actuava de manera autònoma (dècada dels '80 i els primers anys '90).

I llavors va arribar el Web. Per a bé o per a malament. Per a bé, perquè Internet va connectar totes les màquines que produïen (i recaptaven) informació, perquè va abaratir i democratitzar la producció i recol·lecció d'informació i perquè la quantitat d'informació va créixer (y segueix creixent) en proporcions mai imaginades. Però també per a malament, perquè llavors la informació no estava tota en el mateix format, perquè no era necessàriament veritable i perquè podia estar "desactualitzada" o tenir errors.

En aquest context van aparèixer els cercadors d'Internet, per tal d'intentar posar una mica de sentit a tot aquest inabastable oceà d'informació *online*. Els cercadors van evolucionar ràpidament, intentant ajudar, cada vegada millor, a organitzar els milers de milions de documents que es produeixen.

Tipologia. Estructura

A grans trets, hi ha dos tipus de cercadors a Internet: **els directoris i els motors de cerca.**

Els directoris

Són cercadors organitzats a partir d'una jerarquia temàtica (taxonomia). El més conegut dels directoris és Yahoo a la seva pàgina http://es.dir.yahoo.com i Google a les pàgines http://www.google.com/dirhp?hl=ca o http://www.google.com/dirhp?hl=es. Es pot navegar un directori endinsant-nos en les seves categories i subcategories o ingressant una paraula clau que mostrarà les diferents categories i llocs on apareix aquesta paraula.

Hi ha directoris generalistes, com Yahoo!, i especialitzats, com Ariadna, cercador de recursos periodístics (http://www.periodismo.com/buscador/). Tot i que la majoria de cercadors generalistes imiten la taxonomia de Yahoo!, no hi ha un estàndard en aquest sentit. Tampoc hi ha cap tipus d'homogeneïtat o criteri comú entre els cercadors especialitzats. El directori mostra els resultats de la seva cerca basant-se només en el títol i la descripció del lloc. A més, els directoris inclouen llocs complets, no pàgines i seccions dins un lloc. Una altra de les característiques dels directoris és que les pàgines són revisades per

éssers humans i incorporades només si acompleixen els criteris de pertinència del cercador. Això fa que la quantitat de llocs dels directoris sigui petita en comparació amb tots els llocs que existeixen.

L'altre tipus de cercadors són el motors de cerca. El més conegut actualment és Google (www.google.cat). Els motors de cerca no tenen una taxonomia, es pot accedir als resultats només a partir d'una paraula clau. Tot el procés d'indexació de pàgines és automàtic, no hi ha persones revisant cada lloc. A diferència dels directoris, els motors de cerca busquen en tota la pàgina d'un lloc web, no es limiten al títol i a la descripció. Si bé no indexen totes les pàgines d'un lloc, tampoc es limiten a indexar una sola pàgina. La quantitat de pàgines indexades és enorme: en el moment d'escriure això, Google tenia indexades més de 8.000.000.000 pàgines web. Els motors de cerca tenen robots cercadors que exploren els llocs web i els incorporen a les seves bases de dades. Aquesta acció s'anomena indexar.

Actualment molts cercadors combinen la potència dels motors de cerca amb la lògica dels directoris. Si Yahoo! no troba resultats en el seu directori, mostra els resultats del seu motor de cerca. Però també els motors de cerca recorren als directoris: per a qui necessiti resultats més ordenats, Google utilitza les dades del directori Dmoz, conegut també com *Open Directory* (http://dmoz.org/).

Com funcionen

Robots

En una escala microscòpica, els robots dels segle XXI són invisibles i immaterials. Aquests robots es dediquen a fer una cosa clau: indexar les pàgines que es visiten (Piscitelli, 2005).

Avui la situació és molt diferent de fa quatre o cinc anys enrere. En aquell moment, la fe en els robots cercadors feia suposar que si cercadors com Altavista o Hotbot no trobaven allò que buscaven, era senzillament perquè aquesta informació no existia a la Xarxa.

Tanmateix, es va començar a donar importància a la qualitat per sobre de la quantitat. Això va determinar que era preferible indexar llocs de qualitat, abans que apilar la major quantitat de llocs possibles, acudint als cercadors. L'univers Web estava ple de pàgines que no valia la pena de visitar mai.

En aquell moment interessava l'àrea de l'aprenentatge robòtic i es va construir un robot anomenat Inquirus, capaç d'interrogar a altres robots sobre l'existència de documents que complissin certa estructura de cerca; aquest robot podia aportar un benefici secundari més valuós que el que es buscava originalment, estimant la mida real de la xarxa, un número que en aquell moment ningú coneixia amb certesa. Entre els resultats que va aconseguir l'Inquirus aplicat al cercador Hotbot, va descobrir, el 1997, que el Web comptava amb prop de 320 milions de documents (el doble del que es cria abans). I no només això, Hotbot es preava de ser el més exitós i exigent dels robots en aquella època, però, de sobte, es va veure devaluat quan es va descobrir que només indexava el 34% de tota el Web. Com a premi de consol va poder presumir de que als altres

robots els anava encara pitjor: Altavista només cobria un 28% y altres cercadors –com Lycos, que aviat va caure a les mans de Terra i Telefónica– a penes cobrien un 2% de la Xarxa.

El febrer de 1999, quan es va repetir el mateix exercici, els investigadors van trobar que la Xarxa havia crescut (tenia 800 milions de documents), però que la capacitat dels robots d'indexar llocs havia empitjorat.

Un cercador excel·lent de l'època, Northern Light, va ocupar llavors la *pole positon* cobrint el 16% del Web, però Altavista havia baixat al 15% i Hotbot ressenyava a penes l'11% de les pàgines existents. Mentrestant Google, que era un benjamí entre els pesos pesats, a penes veia llavors un 7,8% de les pàgines estimades. El juny de 2001, Google va cobrir per primera vegada 1.000 milions de documents, seguit de prop per Alltheweb. Avui, Google està prop d'arribar als 9.000 milions de documents.

Per molt impressionant que sigui la capacitat d'indexació dels motors, el Web creix infinitament més ràpid que la capacitat que tenen els motors d'analitzar-lo. A més, existeix el Web profund, que és almenys 550 vegades més gran que el que els robots poden arribar, fet pel qual l'asimetria entre el que és visible i el que existeix s'amplia molt més.

L'any 2000, sis de cada deu pàgines no havien estat visitades mai. Avui els resultats arriben a xifres d'entre vuit i nou de cada deu.

Els cercador en el seu rol de gatekeeper

La noció de *gatekeeping* (el porter o vigilant de l'accés) investiga la manera irregular en què les informacions circulen i es troben sotmeses a instàncies que les demoren o traven en algun punt de la cadena de comunicació, i la fluïdesa amb què circulen després les que aconsegueixen passar la barrera. Aquests llocs de demora o nusos que actuen com a barrera i filtre en la circulació de la informació serien eles *gatekeepers* o porters.

El concepte de *gatekeeper* va ser introduït pel psicòleg Kurt Lewin el 1947 mentre treballava en dinàmica de grups i va observar que la informació circulava d'una manera molt irregular, ja que en alguns moments podia interrompre's pels nusos o fluir de manera molt àmplia després de superar-los.

Cal imaginar el cercador com a un *gatekeeper*: en l'univers de totes les les pàgines del Web, el cercador té el poder d'orientar en el camí cap a la cerca de la informació.

Ja se sap que els cercadors no indexen totes les pàgines del Web. Aquí ja hi ha una primera selecció. El *gatekeeper* cercador deixa fora dels seus resultats una gran quantitat de contingut. La segona selecció està en la rellevància: el cercador defineix que determinades pàgines són més importants que d'altres. I es pot comprovar fàcilment que aquest criteri és subjectiu, encara que sigui automàtic, si es comparen els resultats dels diferents cercadors.

Com desafiar aquests criteris? El segon criteri es pot enganyar més fàcilment: utilitzant diversos cercadors i directoris es pot arribar a una "intersubjectivitat de resultats". Hi ha metacercadors com kartoo, turbo10, webcrawler, dogpile,

clusty entre d'altres, per exemple clusty (http://clusty.com/) mostre a l'usuari les millors posicions en què figura cada pàgina en els diferents cercadors.

Desafiar la lògica del cercador pel que fa als continguts que deixa fora dels seus resultats, porta a endinsar-se a l'anomenada Internet invisible.

Com buscar

El primer pas en una cerca és saber què és el que es busca. No necessàriament s'ha de saber amb precisió. Pot interessar, puntualment, trobar la bandera de Rússia o, més vagament, trobar legislació sobre jubilació privada a Amèrica Llatina.

Després de conceptualitzar el que es vol buscar, se sabrà cap on anar. Si la cerca és més específica, es començarà amb un motor de cerca que condueixi cap el lloc que es necessita. Si és més general, serà bo començar per un directori que agrupi tots els llocs comuns al tema que s'investigui.

Si el que es necessita és local o regional, s'haurà de restringir la cerca a aquells països o regions o, millor encara, consultar cercadors de la zona en qüestió.

En aquest sentit, si la temàtica és específica, s'haurà de partir d'un cercador generalista, buscar allà un directori temàtic i realitzar en el directori temàtic una segona cerca, més acotada.

Cada cercador té les seves regles (sintaxi) i per això és recomanable llegir la documentació i les pàgines d'ajuda per entendre bé les seves opcions de cerca.

La majoria dels cercadors accepten els operadors "booleans": AND, OR i AND NOT (aquest últim en alguns cercadors funciona posant només NOT o el signe -).

Per defecte, la majoria de cercadors funcionen amb l'operador AND o +. Això vol dir que posar en un cercador les paraules mapa argentina o mapa AND argentina o mapa + argentina és equivalent.

Posant la paraula OR, mostrarà els documents que continguin almenys una d'aquestes paraules. Per exemple, si es posa argentina OR uruguai mostrarà les pàgines que continguin la paraula argentina, les pàgines que continguin la paraula uruguai i també les que continguin les dues paraules. L'operador OR és també útil si no se sap com s'escriu una paraula (volswagen OR volkswagen), ja que portarà documents que continguin almenys una de les grafies.

L'operador NOT o el signe – exclou paraules de la pàgina de resultats. "Cindy Crawford" – sex – porn –adult –xxx –nude mostrarà documents que mencionin la supermodel, però no contingut pornogràfic. D'això se'n diu filtrar o refinar una cerca. També es pot fer servir si, per exemple, es vol informació només sobre Windows XP però no de Vista haurem de posar windows +XP –Vista).

Les cometes, com en l'exemple anterior, serveixen per indicar una frase exacte: termes que se sap que han d'anar junts, com el títol d'un llibre, d'una pel·lícula,

d'una cançó o d'un joc. S'ha de tenir la certesa que s'escriu de manera correcta, perquè si no ignorarà el que es demana.

Els operadors AND i OR s'han d'escriure en majúscules.

Tots aquests operadors poden ser molt útils si s'utilitzen de manera combinada. Així, per exemple, si volem trobar totes les pàgines en les quals aparegui mencionat Estats Units en català, excloent la grafia en anglès, es pot posar: **"Estats Units" OR EE.UU. OR EEUU -"United States" –USA.**

Cercadors i directoris

- Gigablast Inc. Gigablast <http://www.gigablast.com/>
- Periodismo.com. Ariadna <http://www.periodismo.com/buscador/>
- Google. Google <http://www.google.cat>
- Grub Buscador <http://www.grub.org>
- Ask.com <http://www.ask.com/>·
- IAC Search & Media. Excite <http://www.excite.com/>
- LookSmart, Ltd. Wisenut <http://www.wisenut.com/
- Lycos, Inc. Hotbot <http://www.hotbot.com/>
- Lycos, Inc. Lycos search <http://www.lycos.com/>
- Microsoft. MSN <http://www.msn.com>
- Overture Services, Inc. Alltheweb, find it all <http://www.alltheweb.com/>
- Overture Services, Inc. Altavista <http://www.altavista.com/>
- The New York Times Company. About <http://www.about.com/>
- Walt Disney Internet Group (WDIG). Go.com <http://go.com/>
- WebFile.com. Webfile <http://www.webfile.com/>
- Yahoo! Inc. Yahoo! <http://www.yahoo.com./>

Metacercadors

- ·Copernic Technologies, Inc. Copernic <http://www.copernic.com/>
- ·Digital Tsunami, Inc. Quickfindit <http://www.quickfindit.com/>
- ·Ez2find.com. ez2finf <http://ez2find.com/>
- ·InfoSpace, Inc. Dogpile <http://www.dogpile.com/>
- ·InfoSpace, Inc. Metacrawler <http://www.metacrawler.com/>
- ·Intelliseek, Inc. ProFusion <http://www.profusion.com/index.htm>
- ·Mamma, Inc. Mamma <http://www.mamma.com/>
- ·Surfboard BV. Ixquick <http://www.ixquick.com/>.
- Netscape Communications Corporation. DMOZ Open Directory Project <http://dmoz.org> [

Cercadors de cercadors

- ·Multibuscador.com <http://dir.multibuscador.com/>
- ·Buscopio <http://www.buscopio.net/esp/>

Eines de segona generació: classificació documental

Ens trobem amb un conjunt totalment nou d'eines, diferents a les anteriors perquè són *client-side*. Es tracta, per tant, de programes totalment independents que s'instal·len a l'ordinador client, fet que redunda en un major control i personalització de les seves funcions. El fet que, de vegades, algunes d'aquestes eines poden funcionar de forma autònoma respecte el client en el qual estiguin instal·lades, ha portat a que, incorrectament, es generalitzi el nom d'agent o *bot*, que pot identificar algunes d'elles però no totes.

En general, el conjunt resulta relativament heterogeni, la qual cosa permet construir una classificació molt descriptiva.

A més, com que alguns dels mecanismes són paral·lels als que existeixen com a servidors, aquesta segregació resulta especialment útil i admet anàlisis comparatives de prestacions. Tanmateix, el valor afegit d'alguns d'ells no es restringeix únicament a un increment de la capacitat d'automatització, sinó que ofereixen possibilitats totalment noves. Algunes de les opcions inèdites resulten impossibles d'implementar des d'un servidor.

Entre les novetats més singulars destaquem:

- La possibilitat d'extreure informació d'Internet invisible (infranet), el conjunt de registres de bases de dades o catàlegs de biblioteca accessibles mitjançant formularis web, però que no són indexats pels motors.
- L'ús dels veritables agents que, de manera autònoma, mitjançant mecanismes intel·ligents, poden recórrer la Xarxa, extreure informació i, fins i tot, "aprendre" amb ajuda de l'operador humà. La majoria dels programes revisats són productes comercials disponibles sota el sistema *Shareware* (avaluar abans d'adquirir), cosa que significa que es pot obtenir una còpia d'aquests programes, més o menys operativa, a la xarxa Internet. El preu no és excessivament car, i són precisament els programes més sofisticats els que costen més. Lamentablement, per a aquests tipus de programes a penes s'ofereix suport tècnic i alguns títols, a més, desapareixen ràpidament.

A continuació presentem una classificació comentada d'aquestes eines, utilitzant com a criteri sistematitzador les potencialitats i aplicacions documentals que tenen. Aquest criteri exclou altres programes, relativament nombrosos actualment, de vegades reunits sota la categoria d'"utilitats d'Internet", que són potencialment interessants. L'interès d'aquests programes, sobretot informàtic, pot ser més evident en un futur no gaire llunyà. Segons els usos documentals, distingim cinc grans grups per odre de complexitat:

- Clients Z39.50
- Bolcadors

- Metacercadors
- Indexadors
- Mapadors de port

A banda, també hi ha les eines canalitzadores, que tenen un caràcter mixt. Basades en la tecnologia *push*, podríem qualificar-les d'híbrides, ja que necessiten tant una instal·lació client com un servidor.

La incorporació d'aquest tipus de serveis als clients universals (Netscape y Explorer) ens ha dut finalment a excloure aquestes eines de la nostra classificació, on prèviament les consideràvem "bolcadors" sofisticats. Es pot estar al corrent de les principals novetats d'aquest tipus de programes visitant periòdicament algun dels principals dipòsits de *software* a Internet.

B. El web opac

Pàgines que poden ser indexades, però no són incloses en els cercadors. Els motius perquè els cercadors "decideixen" no incloure-les poden ser:

- Profunditat d'exploració: els llocs tenen "profunditat". La *home page* o pàgina principal és el primer nivell; allà arriben tots els cercadors. Des d'allà s'enllacen a pàgines internes del lloc, aquest seria el segon nivell, al qual no arriben els directoris i alguns motors de cerca. Aquestes pàgines, alhora, enllacen amb pàgines més internes, que no estaven a la pàgina principal. A aquest nivell arriben molt pocs cercadors. Quant més profund sigui el nivell, menys cercadors l'indexaran.
- Freqüència d'exploració: un lloc pot canviar tots els dies, però molts dels robots dels cercadors que exploren els llocs els visiten un cop al mes, o menys, per una qüestió de cost. Tots els canvis entre una visita i una altra no figuren en els cercadors.
- Supera el nombre màxim de resultats: cada cercador defineix quantes pàgines d'un lloc mostrarà. Si un lloc té més pàgines que les que el cercador inclou, les altres quedaran sense indexar.
- Errors d'exploració: pot haver un problema en el lloc o en el robot del cercador (o en la compatibilitat entre els dos) que impedeix que una pàgina (o fins i tot un lloc sencer) s'inclogui en la base de dades del cercador.

Avui dia la majoria dels accessos a llocs web es realitza a través de motors de cerca, per tant, és fonamental per als seus responsables assegurar-se que apareixen ben posicionats en els resultats de cerca, tant des del punt de vista del màrqueting com per donar un millor servei als seus usuaris.

En conseqüència, el lloc que un web ocupa en el llistat de resultats d'un motor de cerca quan un usuari realitza una consulta en el cercador, és un aspecte molt rellevant per als responsables de llocs web, i les accions per a la seva millora s'expliquen pel "posicionament web" (Codina, 2004; Codin, Marcos, 2005; Arbildi, 2005). L'eix central és preguntar-se quines seran les paraules que previsiblement utilitzaran els usuaris potencials en les seves cerques. Una vegada determinat això, es podran utilitzar les tècniques lícites d'optimització perquè un determinat lloc web aparegui en una bona posició quan els usuaris busquin informació relacionada amb els continguts d'aquest lloc web.

Els llocs web que allotgen bases de dades terminològiques també poden i han d'utilitzar les tècniques de millora del posicionament per facilitar als seus possibles usuaris l'accés a aquests recursos a través dels cercadors. La nostra investigació ha pres com a mostra deu bases de dades terminològiques la consulta de les quals és lliure al Web, presenten multilingüisme i pertanyen a diferents temàtiques (taula 1).

Bases de dades	URL
CercaTerm(català)	http://www.termcat.net
Eurodicautom	http://europa.eu.int/eurodicautom
EuskalTerm	http://www1.euskadi.net/euskalterm
OncoTerm	http://www.ugr.es/~oncoterm/
TerminoBanque	http://www.cfwb.be/franca/bd/bd.htm
TIS: Terminological Information System	http://tis.consilium.eu.int
UBTerm	http://www.ub.edu/slc/ubterm
UNTerm	http://unterm.un.org
WTOTerm	http://wtoterm.wto.org

Taula 1. Bases de dades terminològiques estudiades.

Metodologia de la investigació

Per facilitar que els cercadors els trobin, els llocs web han de millorar tots els aspectes que aquests cercadors consideren a l'hora d'establir el rànquing de resultats. Sense entrar a explicar en detall les tècniques per a l'optimització de llocs web, esmentem els criteris que estant usant els cercadors en els seus rànquings de resultats (Codina; Marcos, 2005):

- Freqüència (absoluta o relativa) de l'expressió o terme buscat a la pàgina web (la qual anomenarem "paraula clau"), sempre que no es caigui en una repetició abusiva que els cercadors considerin *spam*. En aquest estudi s'han plantejat tres possibles paraules clau per a cada lloc web.
- Posició: lloc on es troba el terme dins la pàgina; es tindrà en compte que les metadades i el primer paràgraf tenen més pes que altres parts de la pàgina.
- Metadades: a més de les metadades de la secció *head* (principalment *title*, *description* i *keywords*), hi ha altres etiquetes que també proporcionen informació descriptiva útil per als cercadors, per exemple el títol dels enllaços i de les imatges, així com el text alternatiu (alt) de les imatges.
- Popularitat: nombre d'enllaços externs que rep la pàgina web. Aquest criteri està relacionat amb el *PageRank* determinat per la barra d'eines de Google.
- Ancoratge: text que serveix com a enllaç per arribar a aquesta pàgina web.
- Tràfic de visites que rep la pàgina web, considerant alhora el temps de permanència dels usuaris. Ve donat pel *TrafficRank* de la barra d'eines d'Alexa.

D'aquests factors, la nostra investigació està centrada fonamentalment en les metadades i la popularitat. Els altres factors també s'han estudiat, però no han donat resultats que ens ajudin a valorar el grau d'importància que tenen per al posicionament web dels llocs analitzats.

L'estudi ha considerat dos tipus d'anàlisi: en primer lloc, s'ha dut a terme una anàlisi empírico-descriptiva, atenent els aspectes que la bibliografia sobre posicionament web indica que cal tenir en compte. A partir dels resultats obtinguts en aquesta fase, s'han plantejat algunes hipòtesis que s'han posat a prova mitjançant l'anàlisi estadística ANOVA; es tracta d'una anàlisi multivariable en què una variable dependent és creuada amb algunes variables independents.

Per a la seva aplicació s'ha utilitzat el programa Statgraphics Plus 5.0. La variable dependent considerada ha estat el posicionament web dels llocs web, i les variables independents amb les quals s'ha creuat han estat cadascun dels llocs web estudiats, els cercadors sobre els quals s'han fet les consultes, las paraules clau de les consultes i les metadades *title*, *description* i *keywords.* L'encreuament de dades dóna com a resultat el p-valor, que és el nivell de significació empíric en un contrast d'hipòtesis. Es considera que una hipòtesi és nul·la en els casos en què el p-valor és inferior a 0,05, i que és alternativa sempre que supera aquesta xifra.

Elecció de paraules clau per al posicionament web

Abans d'analitzar les causes del posicionament en llocs web, ja que aquesta dada constitueix la variable dependent en l'anàlisi estadístic, s'ha comprovat quin és aquest posicionament per a tres consultes per a les quals aquests llocs web haurien d'estar ben posicionats, entenent per una bona posició els 10 primers resultats dels cercadors més utilitzats avui dia. S'ha valorat de forma especialment positiva (2 punts) que el lloc web aparegui en el primer o el segon lloc, de forma parcial (1 punt) si apareix entre els llocs 3 i 10, i no s'ha valorat (0 punts) si està més enllà del resultat número 10.

Les paraules de cerca escollides s'han traduït a l'idioma principal de la interfície de cada eina terminològica, seguint aquests criteris:

- ➔ Paraula clau 1 (PC1): nom de la base de dades.
- ➔ Paraula clau 2 (PC2): sintagma que descriu el tipus de recurs terminològic de què es tracta en cada cas.
- ➔ Paraula clau 3 (PC3): sintagma que descriu una qualitat específica del recurs terminològic de què es tracta en cada cas.

	PC2	PC3
CercaTerm	Base de dades terminològica	Terminologia catalana
WTOTerm	Terminology database	World trade terminology
TerminoBanque	Banque de données	Terminologie spécialisée
Eurodicautom	Terminology database	Multilingual specialized terminology
EuskalTerm	Banco terminológico	Terminología euskera
TIS	Terminological information system	Consilium terminological database
UBTerm	Base de dades terminològica	Terminologia catalana
UNTerm	Terminology database	United Nations terminology
OncoTerm	Base terminológica de oncología	Terminología oncológica

Taula 2. Paraules clau utilitzades per buscar en els 6 motors.

El mes de gener de 2006 es van realitzar les tres cerques en sis dels cercadors més utilitzats avui dia (Google, Yahoo! Search, MSN Search, Altavista, Teoma i

Vivísimo). El resultat obtingut és molt diferent per a cadascuna de les paraules clau (taula 3):

Les cerques per la PC1 obtenen amb més freqüència el lloc web esperat entre els 10 primers resultats, en especial amb Yahoo! Search i Google, tot i que no tant amb Teoma i Vivísimo, que només aconsegueixen localitzar aquests llocs web en un 50% de les cerques.

El resultat canvia per a la PC2, ja que dos llocs web no s'han trobat entre els 10 primers resultats, i d'altres no ocupen les primeres posicions. Google i Yahoo! Search mostren aquests llocs web en millors posicions que els altres cercadors. Teoma, en canvi, només mostra en aquest primer grup de resultats 4 dels 10 llocs web.

La cerca per la PC3 mostra resultats diferents als anteriors: tot i que segueix en la línia de la PC2 i alguns cercadors no col·loquen aquests llocs web entre els 10 primers resultats, el cercador Teoma aquest cop ofereix les millors posicions, seguit de Vivísimo i Google.

Bases de dades	PC1	PC2	PC3	Mitjana de cada base de dades
Eurodicautom	100,0	50,0	100,0	83,3
TIS	33,0	91,6	100,0	74,8
WTOTerm	100,0	25,0	91,6	72,2
EuskalTerm	100,0	83,3	0,0	61,1
CercaTerm	91,6	0,0	50,0	47,2
OncoTerm	41,6	100,0	0,0	47,2
UNTerm	91,6	25,0	16,6	44,4
UBTerm	75,0	33,0	0,0	36,0
TerminoBanque	100,0	0,0	0,0	33,3
Promedio de cada PC	72,3	50,8	45,8	83,3

Taula 3. Classificació de les bases de dades en funció de la mitjana del posicionament que ocupa cada lloc web al cercador per la PC1, PC2 i PC3. El valor màxim, 100, indica que es troba en el primer o el segon lloc; els valors intermedis, propers a 50, indiquen que es troba entre el lloc 3 i el 10, i el valor 0 indica que no és entre els 10 primers resultats.

L'anàlisi estadística corrobora que existeix una diferència significativa entre el posicionament dels llocs web estudiats i les paraules clau per a les quals s'ha provat el seu posicionament, ja que per a la PC1 el posicionament obtingut és molt millor que per a les PC2 i PC3 (figura 1). Al mateix temps, mostra que l'ús d'un cercador o un altre no provoca diferències significatives en els resultats de posicionament, ja que Google, Yahoo! Search i Vivísimo presenten aquests llocs web més ben posicionats que els altres tres cercadors utilitzats.

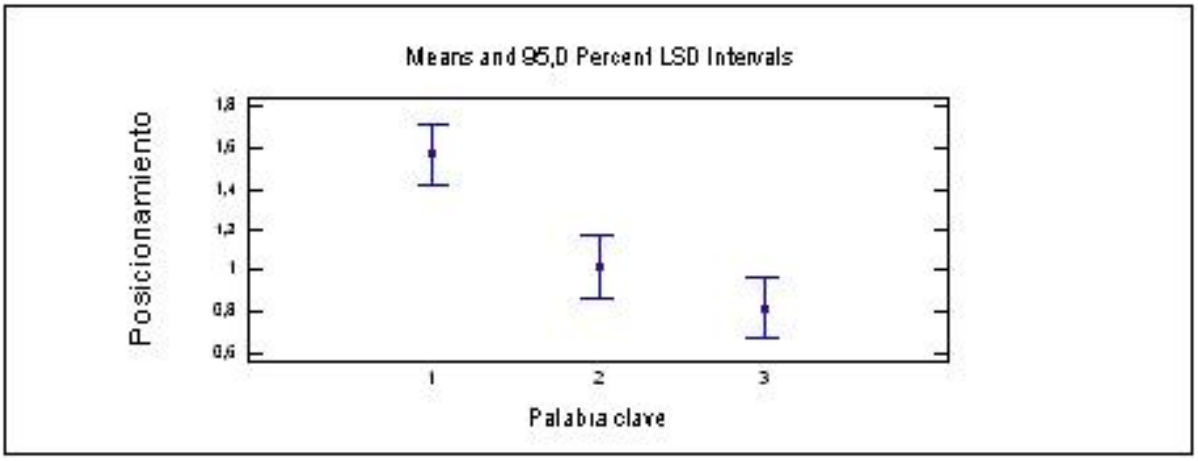


Figura 1. Posicionament dels llocs web en relació amb PC1, PC2 i PC3.

Anàlisi dels factors del posicionament web

A partir dels punts indicats en l'apartat 2, presentem els resultats més rellevants obtinguts en l'anàlisi empírico-descriptiva i en l'anàlisi estadística.

Freqüència d'aparició i posició de les paraules clau

Tot i que estem d'acord amb la importància d'aquests factors per millorar el posicionament web, hem decidit no considerar-ho en aquest estudi per tres motius: en primer lloc, les interfícies de les eines de cerca d'informació terminològica no compten amb un volum de text suficient per poder determinar valors de freqüència òptims; i en segon lloc, amb l'excepció del nom de la base de dades, les altres paraules clau establertes com a consultes (PC2 i PC3) no solen aparèixer reflectides en els sistemes estudiats, és per això que aquest criteri no ens permetrà establir comparacions.

Metadades

Pel que fa a les metadades dels llocs web estudiats, hem comprovat que el 90% tenen el camp *title* emplenat, el 40% inclou el camp *keywords* i només el 20% presenta el camp *description* (taula 3). No s'han considerat les etiquetes *title* i l'*alt* dels enllaços i les imatges, ja que en els llocs estudiats el seu número era molt baix o fins i tot nul, per tant, no afectaria al posicionament web d'aquests llocs.

Bases de dades	Title	Keywords	Description
Base de Terminologie	X	-	-
CercaTerm	X	X	-
Eurodicautom	X	X	-
EuskalTerm	X	X	X
OncoTerm	X	X	-
TerminoBanque	X	-	-
TIS	X	-	X
UBTerm	X	-	-
UNTerm	-	-	-
WTOTerm	X	-	-

Taula 4. Ús de metadades en els llocs web de les bases de dades estudiades.

En les anàlisis estadístiques, l'encreuament del posicionament dels llocs web amb la informació de metadades posa de manifest que només el camp *description* implica diferències significatives de posicionament entre llocs web (figura 2), mentre que l'existència de *keywords* no influeix tant en els resultats obtinguts (figura 3). No s'ha pogut obtenir un resultat fiable sobre la repercussió de l'etiqueta title, ja que 9 de cada 10 llocs web la tenien, és per això que no ens deixa marge per establir comparacions.

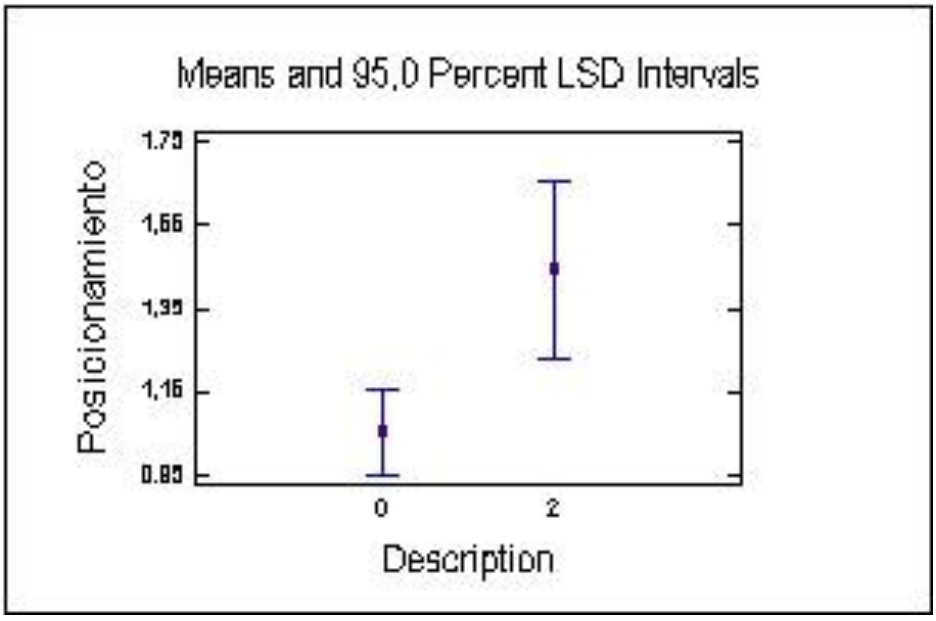


Figura 2. Posicionament dels llocs web considerant si tenen l'etiqueta meta *description* (2) i si no la tenen (0).

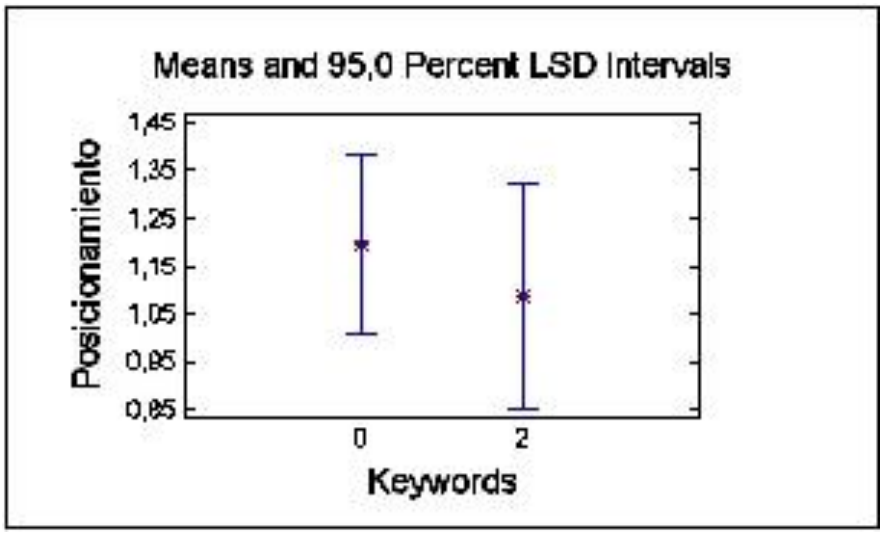


Figura 3. Posicionament dels llocs web considerant si tenen l'etiqueta meta *keywords* (2) i si no la tenen (2).

Popularitat, textos d'ancoratge i tràfic de visites

La popularitat, entesa com el nombre d'enllaços que apunten cap a un lloc web, es pot conèixer parcialment a través de la cerca amb el limitador *link:* que ofereixen alguns cercadors. Mostrem els valors que dóna Yahoo! Search (taula 5), que son xifres molt més altes que les que proporciona Google.

Bases de dades	Enllaços d'arribada (Google)

Eurodicautom	2.340
CercaTerm	1.290
Base de Terminologie	1.030
TIS	458
EuskalTerm	243
TerminoBanque	163
WTOTerm	104
UNTerm	79
UBTerm	49
OncoTerm	9

Taula 5. Classificació de les bases de dades en funció del nombre d'enllaços que apunten a cada lloc web segons Yahoo! Search.

Si comparem els resultats que obtenim en les taules 3 i 5, podem veure com existeix una relació bastant directa entre el nombre d'enllaços d'arribada de cada lloc web i els resultats que han donat per al posicionament per les tres paraules clau. Les bases de dades amb més "cites" són alhora les més ben posicionades per a aquestes paraules. A l'extrem contrari, UBTerm i UNTerm, que obtenen menys "cites", també apareixen en una posició pitjor. El cas de CercaTerm, que no està tan ben posicionada i en canvi rep moltes "cites", és perquè la xifra del nombre d'enllaços d'arribada que té no fa referència a la pàgina principal de la base de dades, sinó a la de la seva institució (TermCat), ja que per arribar a la base de dades cal passar per un formulari a la pàgina d'inici de la institució, cosa que dificulta l'accés directe als cercadors. En el cas d'aquest sistema, si no canvia la manera d'accedir a la base de dades, s'haurà d'optimitzar la pàgina d'inici de la institució per millorar-ne el posicionament en les cerques relacionades amb la seva eina terminològica.

Pel que fa als textos usats en els ancoratges d'aquests enllaços, s'ha revisat els cinc primers de la llista que ofereix Google i el resultat obtingut ha estat que, amb l'excepció de la PC1 (el nom de la base de dades), no s'han utilitzat les paraules clau escollides en aquest estudi. Per tant, és un criteri que no ens serveix per establir comparacions i detectar la rellevància que té en el posicionament de llocs web.

També s'ha anotat el valor que atorga Google como a *PageRank*, tot i que finalment no l'hem considerat d'interès per a aquest estudi, ja que es tracta d'un valor obtingut amb un càlcul poc transparent. El mateix criteri ens ha portat a no

tenir en compte el valor que dóna Alexa com a *TrafficRank*, que, a més, no és específic per a la pàgina web que s'analitza, sinó que agafa el valor de la seva pàgina d'inici.

Conclusions

Tot i que les tècniques d'optimització per a cercadors acostumen a emmarcar-se en el màrqueting, la nostra visió d'aquest concepte és àmplia, ja que creiem que no només es tracta d'una qüestió de publicitat, sinó també de servei als usuaris. En el cas de serveis d'ús gratuït, com és el de les bases de dades terminològiques que estudiem, també haurien de preocupar-se de posar les mesures necessàries per posicionar-se bé, ja que és el primer pas perquè es facin servir.

Quant a la metodologia utilitzada, l'anàlisi empírico-descriptiva ha estat útil per comprovar que cap dels llocs web estudiats ha realitzat una campanya de posicionament. De fet, en la majoria dels casos no s'han emplenat els camps de metadades. Dels camps bàsics que hem mencionat, *title* és el més freqüent (és en el 90% dels llocs web analitzats), seguit de *keywords* (40%) i, molt allunyat, *description* (20%).

No s'ha seguit una política d'enllaços per aconseguir una visibilitat millor, tant per als possible usuaris com per als cercadors. D'altra banda, l'anàlisi estadística multivariable (ANOVA) ha estat vàlid per comprovar les observacions realitzades i verificar hipòtesis. L'aplicació conjunta de les dues metodologies ens du a afirmar que la metadada *description* juga un paper important en el posicionament de llocs web, mentre que la metadada keyword no presenta aquesta evidència (figura 4). De totes maneres, podem afirmar que l'ús de metadades no és decisiu per al posicionament web, perquè es dóna el cas de llocs web ben posicionats que no han fet servir metadades, és per això que no podem deixar de banda altres factors que influencien en el posicionament, com ara la popularitat.

De fet, s'ha pogut observar que existeix una correlació entre la popularitat dels llocs web analitzats (el nombre d'enllaços que apunten cap a ells) i el seu posicionament per a les paraules clau escollides, que indica que el criteri de popularitat juga un paper molt important en la posició que els llocs web ocupen en els resultats de la cerca en motors. Aquesta popularitat hauria d'estar potenciada per uns textos d'ancoratge apropiats, de manera que es consolidin les paraules clau que les institucions escullin per al seu posicionament en el Web. En el cas dels llocs estudiats no s'ha observat aquesta política.

Pel que fa als cercadors utilitzats en l'estudi, Google és el que ha presentat un comportament més regular en relació amb l'ús de les metadades en els llocs web, i Google, Yahoo! Serarch i Vivísimo són els tres motors que han mostrat millors posicionaments per a les paraules clau escollides entre aquests 10 llocs web. En la majoria dels cercadors els llocs web apareixen posicionats quan es realitzen consultes per paraules més específiques (PC1), a excepció de Teoma i Vivísimo, en què els llocs web obtenen millors posicionaments en cerques més generals (PC2 i PC3). Per exemple, en el cas d'Eurodicautom, la cerca per

la PC3 no presenta aquest lloc web entre els primers resultats de Yahoo! Search i, en canvi, per a Teoma ocupa el segon lloc.

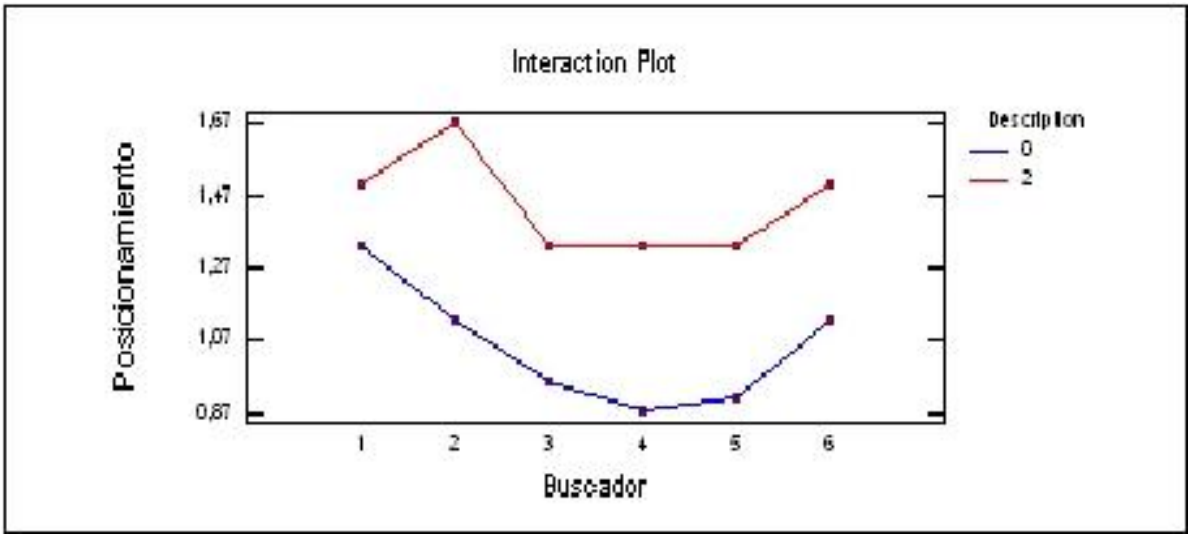


Figura 4. Posiconament dels llocs web en sis cercadors, considerant si tenen l'etiqueta meta description (línia vermella) o no la tenen (línia blava): Google (1), Yahoo! Search (2), MSN Search (3), Altavista (4), Teoma (5) i Vivísimo (6).

Taller pràctic d'indexació

SEO és la tasca d'aconseguir aparèixer en els primers resultats dels cercadors per a determinades cerques, sigla en anglès de ***S**earch **E**ngine **O**ptimization*, optimització per a motors de cerca.

Per aconseguir-ho s'apliquen diferents tècniques i estratègies que poden ser molt diverses. Segons aquestes tècniques, podem classificar SEO en dos grans grups: SEO Black Hat (barret negre) i SEO White Hat (barret blanc).

Tot i que els objectius són els mateixos per a les dues, sortir en les primeres posicions d'un cercador, les seves tècniques són molt diferents. Mentre que el *SEO Black Hat* intentarà amb tots els mitjans aconseguir el posicionament, el *SEO White Hat* també intentarà el mateix, però d'una altra manera més subtil i natural per tal de no posar el *website* en perill. Sobre aquest punt aprofundiré molt més en altres articles.

Avui dia el treball d'un **SEO** és tan ampli i variat que abasta des de sòlids coneixements de programació fins a la psicologia per saber com reaccionaran els usuaris. Per exemple, segons el títol que surti en els *serps* de Google, entraran o no. Un SEO mai deixa d'aprendre o deixarà de ser SEO, ja que els cercadors intenten evitar la seva manipulació dels resultats i corregeixen el seu algoritme segons el que veuen fer als SEO.

Aquí et presentem un petit manual de posicionament web a Google. Amb ell podràs aconseguir aparèixer a les primeres posicions dels resultats. Recorda que només podràs estar a la primera posició si t'esforces molt.

Tenir bons continguts és fonamental per a l'èxit segur d'un lloc web. Per una bada, aconseguiràs atraure un gran nombre de visitants que accediran a les teves pàgines regularment. D'altra banda, si saps com redactar aquests continguts, podràs fins i tot atraure més visites gràcies a Google. Intenta redactar triant determinades paraules clau (keywords) i aprèn on situar-les dins cada pàgina web.

Per què necessito bons continguts?

El contingut és el més important en un lloc web. Podràs conèixer tots els trucs i podràs aconseguir enganyar Google, però la manera d'aconseguir realment visites és amb uns bons continguts.

A més, si els continguts mereixen la pena aconseguiràs més enllaços dels *webmasters* d'altres llocs web. Com veurem més endavant, tenir molts enllaços és fonamental per aconseguir un bon posicionament a Google. No deixis de generar continguts i intenta construir pàgines regularment, amb bona informació.

Haig d'actualitzar constantment els continguts?

És una bona idea actualitzar periòdicament els continguts del teu lloc web per dos motius:

- -A Google li agraden els llocs que renoven i actualitzen els seus continguts. Valora que son llocs "vius" i que es pot comptar amb ells.
- Pots aconseguir que el robot Freshbot passi regularment pel teu lloc web. Aquest robot passa per les pàgines amb els continguts més "frescos" i actualitza aquests continguts a la base de dades de Google en unes hores. D'aquesta manera, pots modificar ràpidament els continguts del teu lloc web (per exemple, amb un nou producte o noves paraules clau) sabent que apareixerà a Google en un parell de dies.

Escull les paraules amb què vols aparèixer a la primera posició dels resultats de Google quan es busca per elles. Per exemple, "coches usados", "abogados en caracas" o "sms gratis".

Planeja amb antelació cada pàgina web i destina dues o tres paraules clau (keyword) per pàgina. És a dir, no intentis que la mateixa pàgina web aparegui en les primeres posicions de Google cercant per moltes paraules. Serà molt difícil aconseguir-ho.

Quines paraules clau (keywords) utilitzo?

Pot ser que tinguis un lloc web dedicat als negocis de cotxes, però no sàpigues quines paraules clau escollir. Una eina molt útil ens la proporciona Google en el KeywordSandbox. En realitat es tracta d'una ajuda per escollir paraules amb el

programa AdWords, però ens pot ajudar molt a l'hora d'escollir les nostres paraules clau.

Per exemple, quan introduïm la paraula "cotxes", aquesta eina ens suggereix "lloguer de cotxes", "cotxes usats" o "cotxes nous", a banda d'altres. A partir d'això, ens haurem de plantejar una estratègia amb les paraules clau.

Altres eines que suggereixen paraules clau:

- http://es.espotting.com/popups/keywordgenbox.asp
- http://inventory.overture.com/d/searchinventory/suggestion/
- http://www.7search.com/scripts/advertiser/sample_get.asp

Tampoc s'han d'oblidar les pàgines que fan una classificació de les paraules més buscades o més populars, com el Zeitgeist de Google. Et podran servir d'ajuda per a noves paraules clau.

Altres pàgines que mostren les paraules més buscades:

- http://sp.ask.com/docs/about/jeevesiq.html
- http://50.lycos.com/
- http://buzz.yahoo.com/

TITLE: probablement el lloc més important. Intenta que en títol de la pàgina web apareguin les paraules clau desitjades. A més, fes un esforç per escriure títols no molt llargs (que no superin els 50 caràcters) i per no repetir més de tres vegades la mateixa paraula (Google ho pot considerar *spam*).

ALT: l'etiqueta ALT està present dins les etiquetes d'imatges d'aquesta manera:

```
<IMG src="mi_imagen.gif" ALT="Mi comentario">.
```

El text de l'etiqueta ALT va sorgir quan hi havia navegadors que no incloïen les imatges, i es mostrava aquest text en comptes de la imatge. Avui dia, alguns navegadors (com MS Internet Explorer) el mostren quan passem el ratolí per damunt la imatge.

Google considera aquest text, sobretot si la imatge és un enllaç a una altra pàgina web. Per això és convenient introduir paraules clau dins l'etiqueta ALT.

METATAGS: Google no considera els continguts dels següents *metatags*:

META NAME=keywords

META NAME=description

Aquest últim, però, és utilitzat de tant en tant per Google en comptes de l'*snippet* (la petita descripció que sol aparèixer en els resultats de Google) quan el contingut META coincideix amb la cerca realitzada.

URL: Google sí que valora que la URL (adreça de la pàgina web) contingui les paraules clau, tot i que no li dóna molta importància. Intenta que contingui les keywords desitjades, però no abusis i no intentis que el domini, el subdomini i el nom de la pàgina continguin aquestes paraules clau. Pots aconseguir que Google et penalitzi.

En les URL intenta separa els noms amb guions "normals" (-) i no amb un guió baix (_). Intenta escriure "mi-pagina.html" i no "mi_pagina.html".

En la resta de la teva pàgina web intenta situar diverses vegades les paraules clau que vols optimitzar. Tampoc abusis d'això, perquè els teus textos seran més difícils de llegir (recorda que dissenyes les pàgines per als usuaris, no per als cercadors).

A més, Google estima que determinats *TAGS* (etiquetes) reflecteixen millor la importància del text. Per exemple, situar un text entre les etiquetes <H1> i </H1> el destaca quan l'usuari veu la pàgina, però també Google considera que aquestes paraules són més importants, i ho tindrà en compte. Passa el mateix amb les etiquetes <H2> (<H3>, <H4>,...), <B> (negreta) i <I> (cursiva). Convé que repassis el vell HTML que has oblidat o que donis un cop d'ull a algun programa d'aprenentatge d'HTML.

Intenta dissenyar les pàgines web i els seus continguts perquè les paraules clau apareguin a les etiquetes, però tampoc abusis d'això, ja que Google ho pot considerar *spam* i et pot penalitzar.

Per altra banda, hi ha eines a Internet que obtenen la densitat de paraules clau del teu lloc web. Pots trobar-les en aquesta cerca:

http://www.google.com/search?q=keyword+density+analyzer.

Cadascuna de les eines et donarà un resultat diferent, perquè en realitat Google utilitza un algoritme bastant complicat per estimar en quin grau una pàgina s'ajusta a determinades paraules clau. De totes maneres, pots utilitzar alguna de les eines suggerides i intentar que la densitat de les teves paraules clau a la teva pàgina web sigui del 5 al 20%.

A més, intenta que les paraules clau que has seleccionat apareguin en els enllaços que apunten cap a les teves pàgines web.

I, evidentment, vigila l'ortografia de les teves paraules clau. Ja sabem que molta gent escriu malament les paraules clau o que es confonen quan escriuen, però si busquem "avogados en caracas", Google ens suggerirà ràpidament "abogados en caracas".

Hauria de tenir el meu propi domini?

Sí. A banda de la millor imatge que pots oferir als teus visitants, pots optimitzar el teu posicionament a Google gràcies als enllaços.

Tenir el teu lloc web a "paginas.sitios-gratis.com/miempresa/" dóna una imatge bastant dolenta, i el preu d'un domini ja no és excusa perquè no en tinguis un de propi. Pots comprar-lo per menys de 10 dòlars l'any. Compara alguns preus dels *registrars* acreditats (és bastant més barat que altres venedors de dominis) i escull el teu.

Quin domini haig d'escollir?

La nostra recomanació és que siguis tu mateix i tinguis la teva pròpia marca. Com es comenta en aquest tutorial, has de dissenyar el teu lloc web per als visitants, no per als cercadors. Si el lloc és bo, la gent recordarà "tunombre.com", però difícilment "abogados-baratos-en-caracas.com".

Fixa't en els exemples de Google i Yahoo!. Són noms de marques no gaire senzills, però han aconseguit que els usuaris les recordin fàcilment. Fins i tot "got.com" va canviar el seu nom per "Overture".

Ara bé, tenir un domini del tipus "abogados-catalunya.com" et dóna l'opció que des d'altres pàgines web t'enllacin així:

<Ahref=http://www.abogados-catalunya.com>abogados-catalunya.com</A>.

Això et donarà la possibilitat d'optimitzar el teu lloc web per a les paraules "abogados catalunya".

A més, si vols aparèixer en els resultats de Google que fan referència a un determinat país, hauràs de tenir un domini del tipus, per exemple, "midominio.cat" (si vols aparèixer en els resultats de Catalunya) o "midominio.cat" (en els d'Espanya). Google també t'inclourà en aquests resultats si el servidor on allotges les teves pàgines web està físicament en aquests països.

Dóna un cop d'ull a la categoria de DMOZ de registre per països. Trobaràs el més adequat per al país en el qual vols aparèixer.

Les pàgines dinàmiques són pàgines HTML generades a partir de llenguatges de programació (*scripts*) que són executades en el mateix servidor web. A diferència d'altres *scripts*, com el JavaScript, que s'executen en el navegador de l'usuari, els "Server Side" *scripts* generen un codi HTML des del mateix servidor web.

Aquest codi HTML pot ser modificat, per exemple, per una petició realitzada per l'usuari en una base de dades. Segons els resultats de la consulta en la base de dades, es generarà un codi HTML o un altre, mostrant diferents continguts.

Quins són els principals tipus de pàgines dinàmiques?

Les pàgines dinàmiques s'executen en el servidor web propi, Per tant, dependran del tipus de servidor que tinguem. Per exemple, si tenim un servidor amb "Microsoft Windows Server", generalment trobarem un servidor web "Internet Inforamtion Server" (IIS) que executarà *scripts* "Active Server Pages"

(ASP). Tot i que això no és sempre així, perquè actualment hi ha paquets de software que executen tots els *scripts* en tots els servidors, sempre estarem condicionats pels llenguatges dissenyats especialment per a cada sistema operatiu.

- **CGI**: abreviatura de Common Gateway Interface. És un estàndard per a la interacció entre aplicacions externes i servidors web. Gràcies a ell, podem adaptar qualsevol programa que hàgim realitzat en qualsevol llenguatge perquè actuï amb el nostre servidor. Tanmateix, Pearl s'ha convertit en el llenguatge més popular per desenvolupar aplicacions CGI, tot i que també se sol utilitzar C, C++ o Fortran.
- PHP: llenguatge *script* de codi obert. Molt utilitzat sobre el servidor web Apache.
- **ASP**: llenguatge *script* creat per Microsoft per al seu servidor web 'Internet Information Server' (IIS), i basat en "Visual Basic Script". L'última versió "ASP.net" forma part del *Framework* ".net".
- **JSP**: llenguatge *script* creat per Sun, basat en la tecnologia Java. No cal que l'usuari disposi de la màquina virtual de Java, ja que es troba en el servidor que crea les pàgines HTML. No té res a veure amb els *applets* de Java, y tampoc amb JavaScript. Els *scripts* JSP són un cas particular dels *servlets*.
- ***Cold Fusion***: llenguatge *script* creat per la companyia Allaire (adquirida més tard per Macromedia). Els *scripts* tenen l'extensió ".cfm".

En què em pot beneficiar utilitzar pàgines dinàmiques?

Les pàgines dinàmiques ens poden ajudar a gestionar més fàcilment els continguts del nostre lloc web i a interactuar am bases de dades.

Per exemple, si tenim un o diversos menús a les nostres pàgines i volem modificar-los, no haurem d'anar pàgina per pàgina per editar-los, sinó que només ho haurem de fer una vegada. En la resta de pàgines només caldrà incloure (en PHP, per exemple): include 'menu-izquierda.html'.

A més, tots els llenguatges *script* comentats tenen components per a la connexió amb la majoria de bases de dades (mySQL, Oracle, SQL Server, etc.). Això ens pot servir per a emmagatzemar els nostres continguts dins d'una base de dades, en lloc de realitzar cada pàgina web una per una.

Informa't de les capacitats de cadascun d'aquests llenguatges *script*, i dóna un cop d'ull als tutorials, que pots trobar a la Xarxa, de CGI Perl, PHP, ASP, JSP i Cold Fusion.

HTML és un estàndard proposat pel Consorci W3C, que té per objectiu aconseguir que tots els documents web siguin compatibles en qualsevol navegador (no només en ordinadors, sinó també en qualsevol dispositiu).

CSS són les sigles de *Cascade StyleSheet*, i especifica la forma del disseny dels documents. Una mateixa pàgina web (un mateix document HTML, per exemple) pot ser vista de diferents formes en un PC que una PDA, gràcies als diferents fulls d'estil CSS.

Utilitzar HTML+CSS et pot ajudar a millorar el teu posicionament web a Google. Per una banda, aconseguiràs que el codi de les teves pàgines web sigui més ampli i clar a ulls del robot de Google. Facilitar la tasca a aquest robot sempre és un punt a favor nostre.

Per altra banda, augmentaràs la densitat de les paraules clau dins els continguts (veieu "on situar les *keywords*"), ja que moltes de les etiquetes t'ocuparan molt menys espai. Això també suposa menys pes per a les teves pàgines web, cosa que Google agrairà. I podràs, alhora, canviar ràpidament els estils d'algunes paraules i modificar així la importància que els hi vols atorgar.

A més, complir amb l'estàndard HTML t'obrirà les portes a dissenyar pàgines web per a dispositius mòbils o per a noves tecnologies que vagin sorgint. I l'ús de CSS et permetrà canviar l'aspecte d'aquestes pàgines web en pocs minuts. En combinació amb les pàgines dinàmiques, pots aconseguir un lloc web realment eficient.

No li ho posis difícil al robot de Google. Si insereixes informació en els següents elements, assegura't que no seran reconeguts:

- ➔ JavaScript
- ➔ DHTML
- ➔ Flash
- ➔ Frames
- ➔ Session IDs
- ➔ Applets de Java
- ➔ Imàtges: no insereixis textos dins d'elles.

Això no significa que no pugis utilitzar-los per al disseny del teu lloc web. Simplement que la informació que aparegui dins d'ells no apareixerà en les cerques de Google.

Per a quin navegador dissenyo el meu lloc web?

No intentis que les teves pàgines web es vegin millor amb un navegador d'Internet determinat o amb un altre. Intenta que es vegin correctament amb tots, però, sobretot, amb el navegador de Google, és a dir, amb el seu robot.

Si vols saber com veu el robot de Google les teves pàgines web, pots utilitzar el navegador Lynx. Es tracta d'un navegador en mode text, que no contempla les imatges ni els elements superflus com el JavaScript, Flash, etc.

Si utilitzes el Sistema Operatiu Linux, probablement el tindràs instal·lat per defecte, i hauràs d'escriure això en la shell:

lynx http://www.mi-sitio-web.com

Si estàs en un altre sistema operatiu (MS Windows, MAC, etc.), el més senzill és accedir via web a un emulador de Lynx:

http://www.delorie.com/web/lynxview.html

La memòria cau de Google ("caché) només emmagatzema un límit de 101 kb. Si mires la informació que guarda Google de qualsevol pàgina, veuràs que no supera aquesta xifra. Se sospita que no s'indexa més enllà d'aquests límits, però no està demostrat.

Aquests 101 kb fan referència només a codi HTML (en el qual s'inclouen tots els textos i la informació). No es tenen en compte les imatges, els gràfics Flash, etc.).

De tota manera, i com a recomanació, intenta que les teves pàgines no tinguin un pes excessiu, és a dir, que ocupin poc espai. Si pot ser menys de 30 kb, millor. Tingues en compte que a la gent no li agrada esperar massa quan accedeix a una pàgina web, i molts es cansen d'esperar després de 3 o 4 segons i marxen a altres pàgines. La connexió per cable o banda ampla no està encara massa estesa i la majoria d'usuaris es connecten a Internet amb un mòdem o compartint la connexió de la seva empresa o la universitat.

PageRank (PR) és un valor numèric que representa la importància que una pàgina web té a Internet. Google considera que quan una pàgina col·loca un enllaç (*link*) a una altra és un vot per aquesta última.

Quants més vots tingui una pàgina, serà considerada més important per Google. A més, la importància de la pàgina que emet el seu vot també determina el pes d'aquest vot. D'aquesta manera, Google calcula la importància d'una pàgina gràcies a tots els vots que rebi, considerant també la importància de cada pàgina que emet el vot.

PageRank (desenvolupat pels fundadors Larry Page i Sergey Brin) és la manera que té Google de decidir la importància d'una pàgina. És una dada valuosa, perquè és un dels factors que determina la posició que tindrà una pàgina en els resultats de la cerca. No és l'únic factor que Google utilitza per classificar les pàgines, però n'és un dels més importants.

Cal tenir en compte que Google no considera tots els *links*. Per exemple, Google filtra i descarta els enllaços de pàgines dedicades exclusivament a col·locar *links* (anomenades *link farms*).

A més, Google admet que una pàgina no pot controlar els enllaços que apunten cap a ella, però sí que pot controlar els que aquesta pàgina col·loca cap altres pàgines. Per això, *links* cap a una pàgina no poden perjudicar-la, però sí que poden ser perjudicials per al seu *PageRankTM* els enllaços que una pàgina col·loqui cap a llocs penalitzats.

Si un lloc web té PRO, generalment és un web penalitzat, i serà poc intel·ligent col·locar un *link* cap a ella.

Una manera de conèixer el *PageRankTM* d'una pàgina és descarregant la barra de cerca de Google (només disponible per a MS Iexplorer). Apareix una barra on es mostra en color verd el valor del *PageRankTM* en una escala de 0 a 10. Llocs web amb PR10 són Yahoo!, Microsoft, Adobe, Macromedia i Google. Teniu una llista completa amb els llocs amb PR10.

L'algoritme de *PageRank* va ser patentat als Estats Units el dia 8 de gener de 1998, per Larry Page. El títol original és *Method for node ranking in a linked database*, i se li va assignar el número de patent 6.285.999.

Aconseguir enllaços és una de les tasques més crítiques en el posicionament web a Google. En funció del nombre d'enllaços que obtinguem, tindrem major PageRank o popularitat.

S'ha de ser extremadament curós amb la manera que realitzem l'enllaç. En funció de com es faci aquest enllaç, promocionarem unes paraules clau o unes altres.

A més, cal saber on aconseguir enllaços. Hi ha determinats llocs web on és més fàcil obtenir-los, però gairebé sempre cal esforçar-se per aconseguir-los i s'ha de fer un seguiment.

Per què cal aconseguir enllaços?

Google atorga un valor numèric a cada pàgina web que insereix a la seva base de dades. A aquest valor numèric l'anomena *PageRank*. Quant més gran sigui el *PageRank* d'una pàgina, més importància li haurà donat Google.

I aquest valor creix quants més llocs web enllacin a la teva pàgina (veieu "transmissió del *PageRank*"). Has de pensar que Google considera cada *link* com un "vot". A més, si aquests llocs web que t'enllacen tenen un *PageRank* elevat, el valor creixerà més, perquè el "vot" és de més qualitat.

Com puc conèixer el PageRank d'una pàgina?

Per conèixer aquest valor, pots descarregar-te la barra de cerca de Google (només disponible per a MS IExplorer), http://toolbar.google.com/ en la qual hi ha un espai per mostrar el *PageRank* (PR) de cada pàgina que visites. Aquest valor varia entre 0 i 10.

Tanmateix, no has de centrar tots els teus esforços en aconseguir un PR elevat. El valor del *PageRank* d'una pàgina és important, però t'hauràs adonat que hi ha pàgines que, tot i tenir menys PR, estan posicionades per sobre d'altres amb un PR més gran per a determinades cerques.

Això ho han aconseguit, entre d'altres coses, perquè han optimitzat millor els continguts de les seves pàgines, perquè han realitzat uns bons enllaços i perquè han aconseguit inserir aquests enllaços en pàgines web bones.

Tot i que Google és el cercador per excel·lència (més del 60% dels usuaris l'utilitzen), no és l'únic. També hauries d'intentar aconseguir tràfic des d'altres cercadors. No tindran un *PageRank* tan gran como Yahoo! o DMOZ, però es tenen en compte tots els enllaços.

Dóna un cop d'ull a les següents categories de DMOZ. Pots trobar alguns directoris i cercadors que et poden ser útils:

1. http://dmoz.org/Computers/Internet/Searching/Directories/
2. http://dmoz.org/Computers/Internet/Searching/Search_Engines/

3. http://dmoz.org/World/Español/Referencia/Buscadores_y_directorios/

Haig d'utilitzar els sistemes automàtics d'enviament a cercadors?

La nostra recomanació és que no utilitzis els sistemes automàtics que prometen que donaran d'alta el teu lloc web a centenars o milers de cercadors. És millor que ho facis personalment, perquè molts cercadors detecten enviaments automàtics i descarten les pàgines web enviades amb aquests mètodes.

A més, aquests sistemes automàtics estan dissenyats per a un determinat moment. La majoria de cercadors modifiquen els formularis i requeriments per donar-s'hi d'alta, i els sistemes automàtics no ho modifiquen alhora. Pot passar també que simplement no funcioni o que, per exemple, no et donis d'alta en la categoria adequada.

Recorda que el posicionament web demana molt esforç i dedicació. Has de tenir paciència i ho has de fer tot personalment.

Com puc aconseguir aparèixer a DMOZ?

Hi ha categories de dmoz.org que tenen un *PageRank* de 7 o 8, això vol dir que aconseguir un enllaç a DMOZ és molt valuós. A més, Google el pren com a referència per construir el seu propi directori "directory.google.cat". No escullis la categoria amb un *PageRank* més alt, sinó la que més s'ajusti a la temàtica del teu lloc web.

Però aparèixer a dmoz.org no és una tasca fàcil, ja que els editors només inclouen llocs web de qualitat i que realment tinguin una relació amb la temàtica de cada categoria.

Navega per les categories de DMOZ. Descobreix quina és la que més s'ajusta a la temàtica del teu lloc web i pitja l'enllaç "agregar URL". Si tens un lloc web en espanyol, el que més et convé és escollir la categoria a "World > Español".

Has de tenir paciència amb la teva sol·licitud. Solen tardar vàries setmanes, però no has d'aclaparar els editors. De totes maneres, pots contactar i debatre amb ells a "www.resource-zone.com". Si en el termini de sis mesos no has aconseguit que t'enllacin, podries tornar a suggerir-los el teu lloc web.

Convé que esperis a tenir el teu lloc web realment llest abans de suggerir un enllaç a la gent de DMOZ. Si els suggereixes un lloc web "en construcció" segur que no te l'accepten i, possiblement, no te'l revisaran la propera vegada que els el suggereixis. No et precipitis i fes les coses amb calma.

Tenen alguna relació Google i DMOZ?

No. DMOZ permet la reproducció lliure dels seus directoris (amb una llicència especial) i Google simplement es limita a recollir els seus continguts des de l'any 2000.

Molts altres llocs fan el mateix que Google en el seu "directory.google.cat", i recullen en les seves pàgines web un directori de categories amb els enllaços

que té DMOZ. Google tracta aquests enllaços de la mateixa manera que la resta dels links, però, en aparèixer a més llocs web –a causa d'aquests "clons" de DMOZ–, el nombre d'enllaços es multiplica.

➔ C. El Web privat / el Web propietari / el Web realment invisible

Diversos especialistes i entitats acadèmiques es dediquen a la tasca d'elaborar i mantenir pàgines concentradores de recursos web seleccionats per àrees d'especialitat (*subject guides*), que poden contenir recursos que no són recuperables amb un cercador comú. Aquests directoris anotats o guies temàtiques solen tenir una gran qualitat, ja que comprometen el prestigi dels autors i de les institucions involucrades. La selecció de recursos sol ser molt acurada i l'actualitzen freqüentment. De vegades, diferents institucions s'associen i formen "circuits" (*web rings*) per a l'elaboració cooperativa d'aquestes guies. Un exemple és *The WWW Virtual Library*.

Els directoris anotats o guies poden incloure, a més, algun mecanisme de cerca en les seves pàgines o al Web en general (Moreno Jiménez, 2004). No n'hi ha prou a conèixer la varietat d'eines de cerca disponibles al Web, sinó que fa falta una orientació sobre el seu funcionament, sobre quines estratègies s'han de seguir per traçar una ruta de cerca adient i sobre com escollir els millors instruments per a cada necessitat. D'això se n'ocupen els tutorials o programes d'aprenentatge. *How to Choose a Search Engine or Directory,* de la Universitat d'Albany, als Estats Units, i les guies de *SearchAbility* i de la Universitat de Leiden, a Holanda, *A Collection of Special Search Engines* orienten l'usuari en l'ampli món tant dels recursos especialitzats en el Web com de les maquinàries que permeten la seva localització.

Però més enllà de totes aquestes eines i recursos es troba el Web invisible.

Sherman i Price identifiquen quatre tipus de continguts invisibles en el Web:

- El Web opac (the opaque web).
- El Web privat (the private web).
- El Web propietari (the proprietary web).
- El Web realment invisible (the truly invisible web).

El Web opac consta d'arxius que podrien estar inclosos en els índexs dels motors de cerca, però no ho estan per alguna d'aquestes raons:

- Extensió de la indexació: per economia, no totes les pàgines d'un lloc són indexades en els cercadors.

- Freqüència de la indexació: Els motors de cerca no tenen capacitat per a indexar totes les pàgines existents; diàriament se n'afegeixen moltes, o es modifiquen o desapareixen i la indexació no es realitza al mateix temps.
- Nombre màxim de resultats visibles: tot i que els motors de cerca presenten un gran nombre de resultats de cerca, generalment limiten el nombre de documents que es mostren (entre 200 i 1.000 documents).
- URL desconnectats: les generacions més recents de cercadors, como Google, presenten els documents per rellevància, basada en el nombre de vegades que apareixen referenciats o lligats a altres documents. Si un document no té un enllaç en un altre serà impossible que la pàgina es descobreixi, ja que no haurà estat indexada.

El Web privat està format per pàgines web que podrien estar indexades en els motors de cerca, però són excloses deliberadament per alguna d'aquestes causes:

- Estan protegides per contrasenyes (*passwords)*.
- Contenen un arxiu "robots.txt" per evitar que les indexin.
- Contenen un camp "noindex" per evitar que el cercador indexi la part corresponent al cos de la pàgina.

El Web propietari inclou aquelles pàgines en les quals és necessari registrar-se per tenir accés al seu contingut, ja sigui de forma gratuïta o pagant. Es diu que almenys el 95% del Web profund conté informació d'accés públic i gratuït (Turner, 2003).

El Web realment invisible està format per pàgines que no poden ser indexades per limitacions tècniques dels cercadors, com ara: informació emmagatzemada en bases de dades relacionals, que no es pot extreure si no es realitza una petició específica. Una altra dificultat és l'estructura i el disseny variables de les bases de dades, així com dels diferents procediments de cerca.

Eines de cerca en el Web profund

Els motors de cerca han millorat el seu funcionament en els darrers anys i permeten un nivell de precisió més alt en les cerques i ofereixen els resultats de manera més útil per a l'usuari. Però encara hi ha molts cercadors que només poden recuperar directament la informació que es troba disponible al Web, però no la informació que s'ofereix a través del Web. Quan es va prendre consciència de la magnitud del Web que resultava invisible per les dificultats que presenten els motors de cerca per accedir a ella, aquests motors de cerca van incorporar funcionalitats addicionals per facilitar la cerca en l'anomenat Web profund. Així, han sorgit cercadors especialitzats en aquest segment del Web. Per afrontar una cerca al Web profund cal tenir en compte que els metacercadors poden presentar limitacions, respecte les possibilitats de cerca

de cada cercador per separat. Per exemple, quan la cerca és sobre materials i formats especials, resulta més fàcil utilitzar les opcions de cerca avançada que tenen el cercadors i, si fos necessari, cal realitzar cerques successives en diversos cercadors o recórrer als directoris concentradors que tenen.

Els mecanisme utilitzats per localitzar recursos en el Web profund consisteixen, sobretot, en directoris de recursos especialitzats, principalment bases de dades disponibles gratuïtament a la Xarxa. El patrocini de les institucions acadèmiques en l'elaboració dels directoris, particularment dels que són anotats, garanteix la cobertura i la qualitat dels recursos compilats. Les guies de recursos especialitzats generalment estan elaborades per bibliotecaris i són una excel·lent erina de cerca i localització de recursos, a més de constituir un bon instrument d'aprenentatge en l'ús de la informació.

Les pàgines *How to Choose a Search Engine or Directory,* de la Universitat d'Albany (Estats Units) i les guies de *SearchAbility* i de la Universitat de Leiden (Holanda) *A Collection of Special Search Engines* inclouen els recursos d'informació i cerca en el web profund.

Finalment, els motors de pregunta dirigida (*directed query engines*) tenen la capacitat de realitzar cerques simultànies en diferents bases de dades al Web. Lexibot i el seu successor, Deep Query Manager, així com Distributed Explorer (Warnick i d'altres) i FeedPoint, són exemples d'aquests motors avançats de cerca.

Estratègies de cerca en el Web profund

A més de les estratègies ja explicades per a la cerca al Web, podem afegir-ne d'altres específiques per a la cerca al Web profund o invisible.

Per a la cerca d'informació especialitzada:

- ➔ Usar les eines de cerca en el Web profund si busquem informació acadèmica de qualitat.
- ➔ Usar cercadors regionals especialitzats per localitzar informació originada fora d'Estats Units o en idiomes diferents a l'anglès.
- ➔ Usar metacercadors per realitzar cerques en diferents cercadors especialitzats alhora.

Per realitzar cerques avançades:

- ➔ Usar les opcions avançades dels cercadors per localitzar imatges o arxius PDF o PostScript.
- ➔ Usar directoris concentradors de cercadors per realitzar cerques avançades successives en uns quants d'ells.

Per avaluar la informació disponible al Web:

- ➔ Usar directoris anotats per avaluar si els recursos disponibles al Web profund són útils per la cerca que estem realitzant.
- ➔ Usar directoris de bases de dades per saber quines d'elles poden oferir-nos informació útil per a la nostra cerca.

Per buscar informació a bases de dades:

- ➔ Usar guies, directoris o motors avançats si la informació que busquem pot estar en una base de dades.

No hi ha dubte que els actuals cercadors i directoris del Web estan millorant el seu funcionament. Més enllà dels detalls tècnics que el públic no pot veure, l'eficiència d'aquestes tecnologies ha augmentat i això s'aprecia en els resultats de les cerques. A mesura que aquestes eines es vagin fent més poderoses, disminuirà la necessitat d'elaborar guies i concentradors de recursos, i també la d'orientar en les estratègies de cerca i en l'ús i aprofitament dels recursos localitzats.

Observant els resultats obtinguts pels motors de cerca, es pot verificar que persisteix encara la pràctica de no indexar totes les pàgines d'un lloc per part dels robots. Per exemple, es pot tenir la referència d'una base de dades que està disponible a través d'un lloc web mitjançant un enllaç a ella que conté una de les pàgines del lloc, en canvi, pot ser que no aparegui la referència a la pàgina d'accés directe a aquesta base de dades en aquest lloc.

És evident que la freqüència de la indexació ha augmentat en alguns cercadors i fins i tot es realitza de forma diferenciada per a alguns recursos. Les pàgines que canvien més (la informació de la borsa, per exemple) serien visitades amb més freqüència pels robots que les que són més estables en el seu contingut.

El nombre màxim de resultats visibles no és un problema quan els cercadors presenten els resultats ordenats per rellevància, ja que sempre apareixeran primer els resultats que s'ajusten més a la cerca realitzada. Quan es pugui realitzar una cerca avançada i els criteris de rellevància combinin el nombre de "lligues" amb la freqüència de paraules, la presentació dels resultats no serà un obstacle per trobar la informació.

L'usuari sempre ha de tenir en compte que els cercadors són més apropiats quan la cerca és específica, és a dir, quan es coneixen dades sobre el que s'està buscant; mentre que és millor realitzar cerques temàtiques en els directoris. Els URL desconnectats podrien evitar-se si existís l'obligació de registrar, encara que fos d'una manera molt senzilla, totes les pàgines que es pengen al Web. Però la gran descentralització d'Internet fa pensar que això no passarà en un futur immediat.

El segment del Web privat no representa una pèrdua de gran valor, pel que fa a la informació que conté, ja que en general es tracta de documents exclosos deliberadament del circuit d'informació per la seva poca utilitat. En qualsevol cal, són els amos de la informació els que decideixen no fer-la disponible, la qual cosa vol dir que difícilment es podran trobar mecanismes legítims per franquejar aquesta barrera. A més, els arxius robots.txt serveixen per evitar que

els robots caiguin en “forats negres”, que els facin entrar en processos circulars interminables, minvant així l’eficiència del seu funcionament.

En un article recent de la OCLC Office for Research (O'Neill; Lavoie i Bennett) s’examinen les tendències pel que fa a la mida, el creixement i la internacionalització del Web públic, és a dir, la part d’informació més visible i accessible per a l’usuari. Les principals conclusions de l’estudi són:

El creixement del Web públic mostra un estancament en els últims anys, perquè es creen menys llocs web i d’altres desapareixen. Però això no vol dir que no augmenti el volum d’informació, és a dir, el nombre de pàgines o el nombre de terabytes. Una altra possibilitat, que no es menciona en aquest estudi però que pot deduir-se de les restriccions per a l’accés a ells, és que alguns llocs web són accessibles mitjançant el pagament d’una subscripció o un altre mitjà de registrament.

El Web públic està dominat per continguts originats als Estats Units d’Amèrica, escrits en anglès. Això ens porta a pensar que probablement hi hagi més recursos invisibles a pàgines originades en d’altres països i en d’altres idiomes.

Alguns cercadors tradicionals com Altavista o Google han evolucionat i presenten ara la possibilitat de realitzar cerques per materials o formats especials. Així, Google ens permet realitzar cerques avançades per localitzar imatges. Per altra banda, el concentrador HotBot presenta la possibilitat de buscar per diferents formats, per localitzar imatges, àudio, vídeo, arxius PDF, Script i Shockwave/Flash. Aquestes opcions estan actives a HotBot per als cercadors Fast (Altheweb) i Inktomi (Pure Web Search), però no funcionen amb Teoma ni Google, tot i que, com vàrem dir, existeix aquesta possibilitat si es realitza la cerca directament des de el lloc de Google.

Aquestes cerques en materials especials, com imatges, àudio i vídeo, són possibles gràcies a una catalogació textual. Les cerques en documents que presenten formats PDF, Flash, etc. es poden realitzar perquè existeixen directoris d’aquests arxius. Així, el principal mitjà amb el qual es poden fer les cerques és el text. Per exemple, si volem recuperar imatges en blanc i negre, aquestes imatges han d’estar classificades d’aquesta manera a la base de dades. Això implica, lògicament, un procés manual. Una pàgina web que conté una imatge, sense cap informació textual sobre el seu contingut, no es podrà recuperar automàticament si no és per la seva extensió (“.jpg”, per exemple).

Com hem vist, la definició més genèrica del que constitueix el Web invisible o profund apunta als recursos que no poden ser recuperats mitjançant les eines comuns de cerca. Per verificar la visibilitat del Web profund, que ha estat identificat pels autors de *The Invisible Web*, Moreno Jiménez (2003) ha seleccionat a l’atzar deu recursos del seu *The Invisible Web Directory* i ha realitzat la cerca en un cercador, un directori, un metacercador i un agent metacercador avançat en la seva versió gratuïta. Els resultats d’aquesta prova senzilla apareixen en el següent quadre:

Recurs	MSN	Yahoo!	MetaCrawler	Copernic
Artcyclopedia	SI	SI	SI (6 cercadors)	SI (8 cercadors)
CRA Forsythe List	SI	SI	SI (3 cercadors)	SI (5 cercadors)
Current Films in the Work (BoxofficeHollywood Hot Set)	SI	SI	SI (3 cercadors)	SI (4 cercadors)
Employee Benefits INFOSOURCE	SI	SI	SI (2 cercadors)	SI (3 cercadors)
Hamnet	SI	SI	SI (4 cercadors)	SI (6 cercadors)
Infonation	SI	SI	SI (5 cercadors)	SI (7 cercadors)
Jourlit	SI	SI	SI (3 cercadors)	SI (7 cercadors)
Scholarly Societies Project	SI	SI	SI (4 cercadors)	SI (6 cercadors)
Vessel Registration Query System	SI	SI	SI (2 cercadors)	SI (6 cercadors)
Who's who in American Art (AskArt)	SI	SI	SI (6 cercadors)	SI (8 cercadors)

Quadre. 15. Resultats de cerca de recursos de *The Invisible Web Directory.*

Tots els recursos seleccionats de *The Invisible Web Directory* són localitzables amb les actuals eines de cerca. A més, en els resultats s'observa que existeixen múltiples referències en altres pàgines, es a dir, que es tracta de pàgines "connectades". L'única dificultat per trobar-les consisteix, en alguns casos, en les paraules amb què es denomina el lloc o el recurs. Per exemple, en el *The Invisible WebDirectory* apareix "Vessel Query Registration System", en lloc de "Vessel Registration Query System", això fa que la cerca per totes les paraules tingui èxit, però la cerca per frase no. Igualment, la denominació de "Who's who in American Art" per al lloc de "AskArt" dificulta la cerca, però si es busca directament pel seu nom apareix en molts cercadors. La taula mostra a més com el solapament entre cercadors és variable.

Es pot dir que el contingut de les bases de dades que estan incloses en aquest directori és invisible, ja que cal realitzar les cerques directament en cadascuna d'elles. Però la veritat és que arribar fins la "porta" d'aquestes bases de dades resulta relativament senzill. El mateix fet que el directori hagi estat col·locat al Web, dóna una visibilitat millor als recursos inclosos, ja que els enllaços en el directori augmenten la possibilitat d'indexació d'aquestes pàgines. Llavors, podem dir que *The Invisible Web Directory* és un bon directori de recursos i

bases de dades disponibles al Web, però no és un bon directori de recursos "invisibles".

En conclusió, el que realment segueix sent invisible al Web són:

- Les pàgines desconnectades.
- Les pàgines no classificades que contenen principalment imatges, àudio o vídeo.
- El contingut de les bases de dades "relacionals".
- El contingut que es genera a temps real.

Però:

- És relativament senzill arribar fins la "porta" de les bases de dades amb contingut important.
- Existeixen motors avançats capaços de realitzar cerques directes simultànies a diferents bases de dades a la vegada; i, tot i que la majoria demanen pagament, també ofereixen versions gratuïtes; el contingut que es genera a temps real per validesa a molta velocitat, tret dels anàlisis històrics.
- És relativament senzill arribar fins la "porta" dels serveis que ofereixen informació a temps real; el contingut que es genera dinàmicament interessa únicament a alguns usuaris amb característiques específiques.
- És relativament senzill arribar fins la "porta" dels serveis que ofereixen contingut generat dinàmicament.

➔ D. El Web realment invisible

Aquest contingut no pot ser indexat pels cercadors per raons tècniques. Els documents poden estar en un format que els robots no reconeixen (música, vídeo, etc.) o perquè són pàgines generades dinàmicament (la pàgina s'autodissenya, no hi ha cap dissenyador humà que la faci o es genera sola). Els cercadors no tenen en compte aquest tipus de pàgines (fòrums de discussió, catàlegs, diccionaris, etc.).

A l'hora d'emprendre una investigació, i després de tenir delimitat inicialment el tema, cal realitzar la pertinent recerca de la documentació bibliogràfica necessària. La ciència és un treball "col·laboratiu" i acumulatiu. Encara que treballem sols, necessitem consultar les teories, models, mètodes, resultats, troballes, dades, etc. aportats per altres autors sobre el mateix tema que hem decidit investigar. Es tracta de no caure en els mateixos errors i/o no duplicar esforços i fer aportacions originals a la comunitat científica. La documentació psicològica és imprescindible en el treball d'investigació científica en Psicologia i es necessita en diverses fases del procés d'elaboració d'escrits científics. En primer lloc, hem de saber identificar la documentació científica rellevant, els

seus tipus i formats (apartat 2). Referent a això, la situació actual és la d'una imparable migració de les fonts documentals en Psicologia de publicacions impreses (format paper) a publicacions electròniques, principalment a Internet. Després entraríem en la documentació bibliogràfica en Psicologia (apartat 3), tant en format imprès tradicional, com a Internet, ja sigui "formal" o "informal" o altres informacions disponibles a Internet.

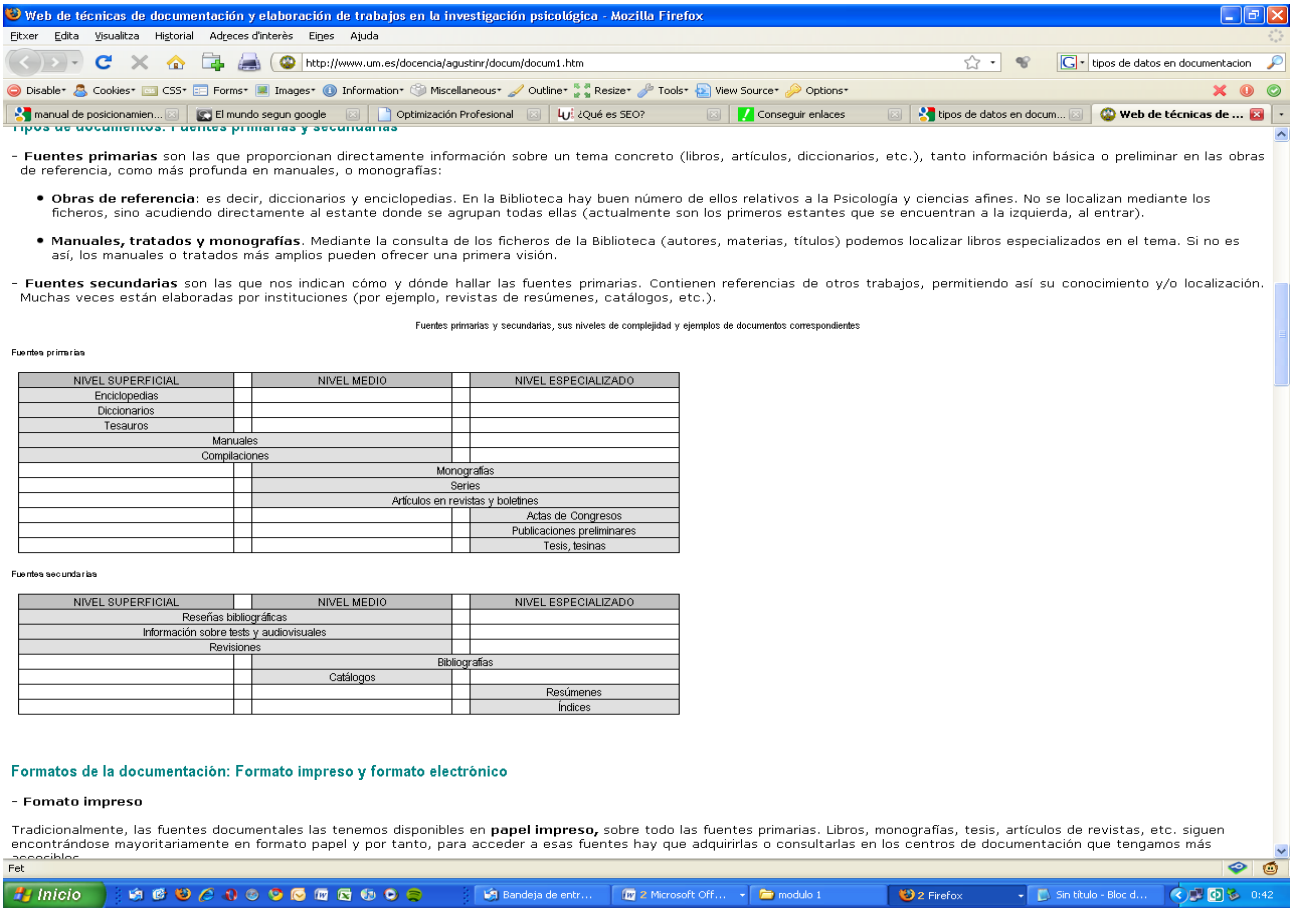
Què buscar: tipus i formats de la documentació bibliogràfica

- Tipus de documents: fonts primàries i secundàries.
- Formats de la documentació: imprès i electrònic.

Tipus de documents: fonts primàries i secundàries

- Fonts primàries: són les que proporcionen directament informació sobre un tema concret (llibres, articles, diccionaris, etc.), tant informació bàsica o preliminar en les obres de referència, com més profunda en manuals o monografies:
- Obres de referència: és a dir, diccionaris i enciclopèdies. A la biblioteca hi ha un gran nombre d'elles relatives a la Psicologia i ciències afins. No es localitzen mitjançant els fitxers, sinó acudint directament al prestatge on s'agrupen (actualment són els primers prestatges que es troben a l'esquerra, a l'entrar).
- Manuals, tractats i monografies. Mitjançant la consulta dels fitxers de la Biblioteca (autors, matèries, títols) podem localitzar llibres especialitzats en el tema. Si no és així, els manuals o tractats més amplis poden oferir una primera visió.
- Fonts secundàries: són les que ens indiquen com i on trobar les fonts primàries. Contenen referències d'altres treballs, permetent així el seu coneixement i/o localització. Moltes vegades estan elaborades per institucions (per exemple, revistes de resums, catàlegs, etc.).

Fonts primàries i secundàries, els seus nivells de complexitat i exemples de documents corresponents.



Formats de la documentació: format imprès i format electrònic

Fomat imprès:

Tradicionalment, les fonts documentals les tenim disponibles en paper imprès, sobretot les fonts primàries. Llibres, monografies, tesis, articles de revistes, etc. segueixen trobant-se majoritàriament en format paper i, per tant, per accedir a aquestes fonts cal adquirir-les o consultar-les en els centres de documentació que tinguem més accessibles.

Format electrònic:

Cada vegada són més útils i es disposa de més fonts en format electrònic, tant fonts primàries (revistes electròniques accessibles per Internet) com sobretot fonts secundàries organitzades com bases de dades referents a catàlegs de biblioteques, índexs de sumaris o resums de revistes, etc.

El format electrònic fa referència a suports magnètics accessibles directament (discos flexibles, discos durs, discos compactes per a CDROM, etc.) o via telemàtica (Internet, etc.) que contenen bases de dades en les quals s'emmagatzema la informació i es processa i recupera per mitjans informàtics. Tant fonts primàries com secundàries es troben ja en aquest suport, més les

segones que no pas les primeres. Així, cada vegada tenim més articles de revistes en format pdf o html a Internet. Pel que fa a fonts secundàries, el més utilitzat són les bases de dades sobre:

- Resums i referències d'articles de revistes i capítols de llibres.
- Llibres.
- Resums de tesis doctorals i memòries de llicenciatura.
- Disposicions legals.
- Catàlegs comercials d'editorials i empreses de *software*, etc.

Avui dia, en general, les fonts secundàries importants per a la investigació psicològica estan accessibles a les universitats des de qualsevol ordinador connectat a Internet.

Què buscar a Internet:

- Documentació bibliogràfica, tant fonts primàries com secundàries.
- Informació temàtica "informal".
- Informació sobre centres i recursos.
- Intercanvi d'informació sobre temes concrets.

Documentació bibliogràfica, tant fonts primàries com secundàries

Fonts primàries

- Podem trobar articles de revistes i fins i tot llibres al Web. El problema és la dificultat per trobar-los, ja que de vegades es tracta d'un autor que al web del seu Departament, en un apartat de professors, i en la seva pàgina particular ha posat la seva bibliografia més recent.
- Revistes electròniques: es tracta de revistes que han sorgit en el Web, o bé s'editaven normalment en paper i ara també han passat a format electrònic (accessibles gratuïtament des de la Xarxa).

Fonts secundàries

Bases de dades de centres espanyols accessibles des del Web i amb documentació de diverses àrees. Cal distingir dos tipus:

1) Bases de dades comercials subscrites per les universitats: es tracta de la documentació que només es pot aconseguir mitjançant pagament o

subscripció, sigui de manera personal o per una institució o centre de documentació (que si és pública permetrà l'accés als investigadors). Aquesta informació sol estar emmagatzemada en bases de dades elaborades per empreses de documentació i que cobren per accedir a elles. Les universitats les subscriuen amb llicència de xarxa i des dels seus ordinadors s'accedeix lliurement a elles (no des d'ordinadors externs a la universitat).

2) Fonts secundàries de domini públic al Web:

- ➔ Catàlegs de biblioteques d'universitats i centres d'investigació.
- ➔ Índexs de revistes: algunes revistes que s'editen en paper ja exposen els seus índexs i fins i tot resums al Web.
- ➔ Bases de dades d'índexs o resums de revistes.

Informació temàtica "informal":

Quant a les cerques per àrees temàtiques, sobretot en directoris temàtics, el seu avantatge, i també el seu inconvenient, és que les adreces web que inclouen estan preseleccionades per l'autor o autors del web, és a dir, depenem dels seus criteris de selecció, normalment esbiaixats en funció de la seva orientació empresarial, teòrica (en psicologia, orientacions cognitiva, psicodinámica, etc.) o professional.

Informació sobre centres i recursos:

És la informació típica, la que primer va anar sorgint al Web. Els centres o institucions que tenien servidor web, el primer que van fer va ser penjar informació institucional sobre ells.

Intercanvi d'informació sobre temes concrets:

El sistema més clàssic és el de les *News* (ja esmentat). També s'han utilitzat llistes de correu, dins el sistema de correu electrònic, o accessibles des del Web. També el sistema xat o IRC (*Internet Relay Xat*).

Com buscar els documents en format imprès

- ➔ Biblioteca i/o hemeroteca de Facultat o d'Universitat.
- ➔ Biblioteca d'àrea o de departament universitaris.

Altres:

- ➔ Centre de Documentació del Col·legi Oficial de Psicòlegs de Madrid.
- ➔ CINDOC (Centre d'Informació i Documentació Científica).

Com buscar els documents a Internet

Els cercadors o motors de cerca:

L'estratègia de cerca d'informació més comuna és a través de cercadors o motors de cerca. Es tracta de webs que generen llistats d'adreces web després d'introduir les paraules clau específiques del tema buscat. El més potent és Google en l'adreça http://www.google.cat.

Aquí podem trobar una àmplia varietat de temes. Tanta varietat i tants nodes web hi ha sobre tot això, que es corre el perill de perdre una gran quantitat de temps buscant el que interessa. Per aquest motiu és fonamental conèixer els dos sistemes de cerca disponibles:

- ➔ A través de motors de cerca (*search engine*) al Web, tipus Google, Yahoo, Lycos, etc. Són sistemes que funcionen amb paraules clau; el resultat de la cerca és un llistat d'adreces web en els quals s'esmenten temes relacionats amb les paraules clau buscades. Avui dia pràcticament el més útil, actualitzat i ampli és Google.
- ➔ A través de directoris elaborats per institucions o persones sobre temes o aspectes concrets.

Els avantatges i inconvenients d'aquests sistemes són:

L'avantatge dels cercadors generals per paraules clau radica que les seves bases de dades estan permanentment actualitzades, inclouen adreces web de manera automàtica –l'autor del web no ha de sol·licitar la seva inclusió, un o dos mesos després de publicar una web ja apareix en algun d'aquests cercadors– i a més, el temps de cerca és molt breu (menys d'un segon per al programa de cerca i uns pocs segons més a causa de la velocitat de la nostra connexió a Internet). Ha d'utilitzar-se aquest recurs quan vulguem trobar agluna cosa específica, amb rapidesa i en gran quantitat.

L'inconvenient és que cal afinar molt en les paraules clau i en el llistat que ens retorna el cercador solen entrar moltes adreces web inservibles (el que es diu "soroll" a Internet). És a dir, els motors de cerca recullen pàgines web automàticament i, per tant, sense intervenció humana i sense control de qualitat.

Un altre problema és que de vegades les recopilacions d'hiperenllaços no estan molt actualitzades (actualitzar les adreces web és una tasca molt laboriosa i a més cal comprovar periòdicament si els enllaços funcionen, etc.).

Criteris de qualitat de la informació de les pàgines web

En aquests casos, el problema és la sobreabundància d'informació i saber discriminar la informació de qualitat. Per a això, a *Romero* (2002) s'especifiquen una sèrie de criteris que seria convenient tenir en compte per determinar la qualitat d'un web sobre psicologia:

1) El primer és excloure de la recopilació:

- ➔ Pàgines primàriament comercials.
- ➔ Pàgines basades en investigació escassa o poc fonamentada.
- ➔ Absència d'un patrocinador clarament identificat.
- ➔ Absència de continguts suficients sobre el tema.

2) Excloses aquestes pàgines web, s'han d'utilitzar les que se'ls pugui assignar un mínim de dos punts (sobre una escala de 5) en els següents criteris:

- ➔ Continguts, amb objectivitat, originalitat i citació de les fonts de les troballes d'investigació i estadístiques.
- ➔ Autoritat, basada en les credencials, tant de l'organització que patrocina com dels autors individuals de la informació presentada. Les credencials inclouen factors com l'estatus educatiu de directors –*staff* i autors–, nombre i qualitat de les publicacions d'investigació, afiliacions institucionals, experiència professional, etc.
- ➔ Actualització i estabilitat, presència de data de creació o *copyright*, evidència de manteniment del web, com dates d'actualització del web o consistència de dates en pàgines interiors.
- ➔ Facilitat d'ús: accessibilitat del material en el lloc web, facilitat de navegació, format consistent i coherent de totes les pàgines del lloc web, operativitat dels enllaços, temps de descàrrega del web acceptable.

En una puntuació global del lloc web avaluat, els dos primers criteris tenen més importància que els dos últims (en els criteris abans mencionats, cadascun dels dos criteris suposarien el 36% de la puntuació global, i els altres dos un 14% cadascun).

3) Per altra banda, a l'hora de descriure els llocs web de psicologia, a més dels criteris anteriors, hauríem de tenir en compte:

- País d'origen, idioma, enllaços en espanyol i/o en anglès.
- Temàtica.
- Si contenen textos complets (fonts primàries) i/o referències (fonts secundàries).
- L'enfocament o utilitat de la informació (pràctica, divulgació, professional, ensenyament/docència, investigació, etc.).

➔ E. Internet invisible

Definició i reptes

Lluís Codina és professor titular de Ciències de la Documentació a la Universitat Pompeu Fabra i membre de l'Observatori de la Comunicació Científica. És sense dubte uns dels millors experts en Anàlisi i Mètodes en Ciències de la Documentació. Podeu trobar molta documentació a la seva web http://www.lluiscodina.com/. Lluis Codina afirma que:

"Internet invisible és un nom clarament inadequat per referir-se al sector de llocs i de pàgines web que no poden indexar els motors de cerca d'ús públic com Google o Altavista. Malgrat el nom, afortunadament, el Web invisible és perfectament visible ja que els continguts d'aquestes pàgines i llocs web es poden veure o bé mitjançant un navegador convencional o bé mitjançant un navegador complementat amb algun programa addicional (*plugin*). Per aquest motiu, s'hauria de denominar, en realitat, el web "no indexable", un terme molt més adequat, però clarament allunyat de la capacitat suggeridora del terme "invisible". Atès que, tot i així, és el terme més habitual fins i tot en la bibliografia tècnica, usarem en aquest treball el terme Web o Internet invisible per referir-nos a la informació publicada en servidors web que per diversos motius no pot ser indexada i, per tant, no pot ser trobada pels motors de cerca convencionals."

Vegem ara per què hi ha continguts no indexables al Web. Hi ha almenys tres motius. En un ordre no significatiu, podem dir que el primer motiu són els formats dels documents. Els motors de cerca van ser creats originalment per descarregar, llegir i indexar pàgines HTML. Qualsevol altre format era il·legible, és a dir, invisible per a aquests motors. Tots coneixem la proliferació de formats no HTML en el Web (que no obstant això s'integren amb tota facilitat en el navegador). És el cas, per exemple, dels cada vegada més abundants documents en format .pdf (documents Acrobat) i fins i tot en format .doc (documents Word). En la mesura que una part dels continguts del Web està formada per documents no HTML, aquesta part és candidata a ser Internet invisible.

Figura 1. Part d'un document –un dibuix d'un tomàquet– en format no HTML (SVG) vist en un navegador.

El format de gràfics vectorials SVG, Scalable Vectorial Graphics, és un format estàndard per fer dibuixos, plànols, esquemes, etc, per pàgines web, en aquest cas el dibuix d'un tomàquet molt realista. SVG està basat en XML i per tant és indexable pels cercadors. El format Flash (animacions) ja és indexat també per Google. Més informació a http://es.wikipedia.org/wiki/.svg.

El segon motiu són les pàgines que es generen de forma dinàmica; típicament, a través de la consulta a una base de dades. Per exemple, si utilitzem All Movie (www.allmovie.com) per buscar informació sobre un film obtindrem una URL com aquesta:

http://www.allmovie.com/search/work/star+trek/results

Els motors de recerca no poden indexar continguts que es generen d'aquesta manera. Abans de llançar la cerca, el contingut existeix en el format binari (i propietari) d'alguna base de dades. Solament després de la consulta, i com a resultat d'executar una instrucció com la que mostra la figura anterior, es crearà una pàgina en format HTML. El lector pot fer la prova, si copia la URL de la figura anterior (que conté una consulta a una base de dades) i la introdueix com a adreça en un navegador, obtindrà una pàgina HTML que li informarà sobre un film determinat. Abans, però, aquesta pàgina no existia. En el cas de bases de dades com l'anterior, els motors de cerca poden proporcionar accés a la pàgina d'inici (*home page*).

És a dir, podem accedir a les pàgines principals dels llocs web que proporcionen accés a bases de dades, perquè aquestes pàgines principals són

pàgines HTML convencionals, però no podem accedir a la resta del lloc a través del motor de cerca; i la resta del lloc pot ser (en ocasions) una enorme base de dades. Per exemple, si llancem la consulta 2001 a Google, en cap dels resultats obtenim la fitxa del film corresponent d'All Movie. De fet, obtindrem una diversitat de resultats que reflecteix que el terme 2001, fora de context, té molts significats i no necessàriament el de títol principal d'un film de Kubrick.

Finalment, forma part del Web invisible el conjunt de llocs o de pàgines web que, de forma expressa, s'exclouen de l'activitat indicadora dels motors de cerca. Alguns servidors exclouen als motors de recerca de totes o de part de les seves carpetes i directoris mitjançant l'ús d'un protocol d'exclusió que, en general, respecten els programes rastrejadors (*spiders* o *crawlers*) d'aquests motors de recerca. Aquest protocol consisteix en un petit nombre de valors que pot adquirir l'atribut *content* com a part d'una etiqueta meta i l'altre atribut, *name*, obté el valor "robots". Aquestes indicacions es guarden en un simple arxiu de text de nom robots.txt que se situa en el servidor de pàgina web i que se suposa que llegeixen i respecten els rastrejadors (robots). La figura següent mostra l'ús d'aquest protocol per a indicar als robots dels motors que no indexin la pàgina en qüestió ni segueixin cap dels enllaços que pugui contenir.

```
<meta name="ROBOTS" content="noindex,nofollow">
```

A més del protocol que acabem de veure, hi ha altres raons per les quals els motors no poden entrar en un lloc. En general, qualsevol lloc web que demani l'ús de contrasenyes o *passwords* quedarà fora de la capacitat indexadora dels motors. Aquests llocs poden ser Extranets o serveis que demanin, no només una subscripció prèvia, sinó que el pagament d'una quantitat en concepte d'abonament, etc. Els motors també tenen dificultats per interpretar els llocs que usen marcs (*frames*), encara que són dificultats d'un altre tipus i no les considerarem aquí.

La qüestió és que, en total, alguns analistes assenyalen que el Web invisible pot ser fins a 500 vegades més gran que el Web visible (Bergman, 2001). Des del punt de vista de l'accés al coneixement i de la classe de cerca i obtenció de la informació que ens interessa aquí, no hi ha cap problema amb que una part del Web invisible segueixi sent invisible.

Per exemple, no és cap tragèdia per al desenvolupament de la ciència o del coneixement humà que l'Extranet o la Intranet d'una corporació sigui invisible als motors de recerca. No només no és un problema, sinó que és desitjable que segueixi sent així. Ningú vol que els motors de cerca puguin indexar documents administratius particulars o informacions confidencials.

Per tant, de les tres raons per les quals tenim una Internet invisible, una d'elles no és cap problema, però les altres dues sí. Recordem: documents amb format no HTML i pàgines generades dinàmicament (típicament a través de bases de dades).

Amb la impossibilitat d'indexar documents no HTML tenim, efectivament, un autèntic problema. Molts informes i estudis que contenen informació valuosa estan publicats i disponibles al Web de forma pública i oberta; tanmateix, si no són indexades de forma adequada, són inaccessibles a gairebé tot el món i a tots els efectes pràctics.

D'altra banda, no deixa de ser un problema que, tot i disposar d'un client universal d'accés a la informació –el navegador web–, no existeixi, en canvi, alguna cosa similar a una interfície universal d'accés a la informació des del moment que, per a cadascuna de les vàries desenes de milers de bases de dades existents a Internet sigui necessari: primer, un accés diferenciat, i segon, un sistema de consulta (en part) diferent.

En aquest últim cas, cal observar que les barreres al coneixement són dues: el coneixement de les fonts i el domini de la interfície d'usuari de cada font. En efecte, en primer lloc, perquè un usuari pugui beneficiar-se dels continguts d'una base de dades és necessari, almenys, que sàpiga que existeix. Però, suposant que sàpiga que existeix, llavors haurà de tenir habilitats d'ús d'aquesta base de dades, i cada base de dades, no només presenta una interfície d'usuari diferent, sinó un conjunt de funcions diferents.

Accedir als continguts d'Internet invisible

Formats no html

Malgrat tot això, es pot accedir cada vegada més a parts més grans del Web invisible. Examinem primer el cas dels formats de documents. Afortunadament, en aquest aspecte, les fronteres del Web invisible no fan més que retrocedir.

Google té capacitat per a localitzar una gran varietat de documents en diferents formats (pdf, excel, word, access, flash, rtf, postscript, i molts més). L'últim format incorporat més destacable és el dels arxius swf confeccionats en Flash.

En aquest sentit, sembla que la tendència és clara: a poc a poc, la major part dels formats de documents significatius en el món científic i cultural seran indexats pels motors de cerca i, per tant, aquesta zona del Web invisible deixarà de ser-ho aviat. A més, hi ha dos factors més que conflueixen en aquest aspecte: d'una banda, els navegadors cada vegada incorporen amb major facilitat documents no HTML. És exemplar, en aquest sentit, la integració de les últimes versions dels navegadors i el format pdf. D'altra banda, el progressiu ample de banda disponible per als usuaris fa que aquesta integració sigui transparent.

D'aquesta manera, si els motors tendeixen al que podríem anomenar una "indexació universal" i els navegadors (o agents d'usuari) tendeixen a poder mostrar qualsevol tipus de document, podem concloure que aquest aspecte del Web invisible està destinat a ser marginal.

Ara bé, de vegades les solucions als problemes aporten també problemes nous. A mesura que formats com pdf i Word s'integren al Web amb major naturalitat, per a benefici dels usuaris, descendeix el grau de connectivitat general del Web. És a dir, una de les virtuts del Web és la facilitat amb la qual

es poden publicar pàgines web (o llocs sencers) interconnectats de forma interna, així com la facilitat per connectar pàgines i llocs web remots. Però part d'aquestes facilitats desapareixen amb formats com pdf i Word. És cert que un document pdf, per exemple, pot contenir enllaços interns o externs, però en la pràctica, es publiquen documents pdf com una forma fàcil d'obtenir una publicació de qualitat tipogràfica amb un mínim esforç. En la pràctica, per tant, la immensa majoria de documents pdf estan interconnectats de manera molt pobre.

Bases de dades

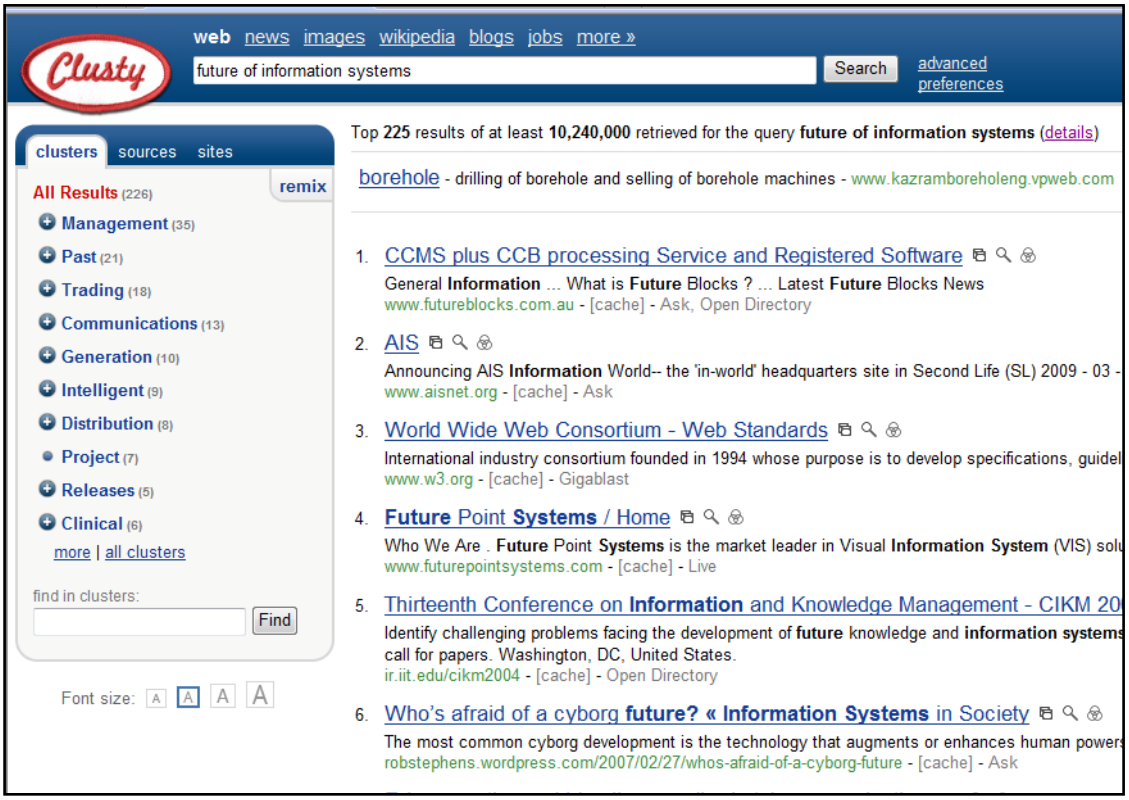
També tenim indicis de solució al segon gran "problema" del Web invisible: l'accés al contingut de les bases de dades, però des de motors convencionals.

La solució aquí prové d'aquest enfocament: si bé és difícil o impossible indexar per part dels motors de cerca el contingut de bases de dades alienes, no hauria d'haver molta dificultat a generar interfícies de consulta unificades que enviessin una mateixa consulta a diferents bases de dades des de, per exemple, una mateixa pàgina web. El model en aquest cas són els multicercadors, també (mal) anomentas metcercadors.

Un multicercador és un sistema que accepta com a entrada la pregunta d'un usuari i retorna en una resposta unificada les respostes de diversos motors de cerca.

Un bon exemple de multicercador és http://clusty.com. Una cerca en Clusty pels tremes *future of information systems* mostra com a resultat una compilació de la informació oferida per diversos buscadors.

Figura 8: El resultat d'una cerca en Clusty:



Compilar informació, en el cas de Clusty, significa que no es limita a bolcar els resultats que envia cada cercador, sinó que: (a) unifica resultats (o sigui, elimina duplicats); i (b) distribueix els resultats per grups o pseudocategories que el sistema d'agrupació (*clustering*) de que és capaç de generar de manera automàtica.

Però el que ens interessa aquí examinar és la següent idea: Clusty no intenta explotar directament els índexs dels diferents motors de cerca. En el seu lloc, fa una cosa més viable: envia la pregunta a diversos motors i processa els resultats abans d'oferir-los a l'usuari. Aquesta operació li permet oferir un resultat unificat les fonts del qual, però, tenen procedències molt diverses.

Multicercadors de segona generació

Un altre exemple molt interessant i bona mostra del que, probablement, ens espera en els propers anys és el motor de cerca Scirus (www.scirus.com). És aviat encara per saber si Scirus serà un experiment efímer, com tants altres projectes esperançadors en el Web (esperem que aquesta vegada no) o solament un avançament d'una nova generació de sistemes de cerca en línia que trenqui d'una vegada per sempre les barreres del Web invisible.

Scirus és un projecte d'una important editorial científica, Elsevier, que ha produït un motor que és capaç d'enviar les preguntes dels usuaris a les bases de dades que indica la taula de la Figura 10.

Figura 10. Bases de dades que pot interrogar Scirus de forma simultània:

- Medline
- Sciencedirect
- Uspto
- Beilstein Abstracts
- E-Print Arxiv
- Nasa Technical Reports
- Cogprints
- Biomed Central
- Mathematics Preprint Server
- Chemistry Preprint Server
- Computer Science Preprint Server

A més, Scirus indexa gairebé 90 milions de pàgines web, és a dir, documents en format HTML publicats en servidors de pàgines web convencionals, però sempre vinculats amb institucions acadèmiques o científiques. D'aquesta manera, l'usuari de Scirus, típicament un investigador o un professional, quan realitza una cerca en aquest motor, obté dos tipus de resultats: (1) pàgines o llocs web relacionats amb la ciència, la universitat, etc.; (2) articles de revista o registres referenciales procedents de bases de dades de ciència i tecnologia (és a dir, una part del Web invisible).

Scirus, per tant, és un dels millors exemples que tenim ara al nostre abast del que poden ser els futurs sistemes d'informació en línia: una interfície unificada d'informació a fonts diverses.

Podem concloure, en relació a aquest apartat, que les barreres d'Internet invisible probablement cediran, una a una, fins que els continguts no indexables d'Internet siguin exactament els que han de ser: porcions del web que els seus administradors o propietaris, en ús legítim de les seves prerrogatives, no desitgen que siguin indexats.

En canvi, els continguts de la Internet invisible corresponents a formats no HTML i part del contingut que es troba en el format binari de diferents bases de dades, seran accessibles des de motors de cerca públics, del tipus Google o Scirus.

El que significa això últim és que els productors de bases de dades haurien de començar a plantejar-se si desitgen, per dir-ho d'alguna manera, sindicar els seus continguts als motors de cerca. Un model pot ser el que representa Scirus. Els productors de bases de dades poden decidir que entra en els seus interessos permetre la recepció de consultes i l'enviament consegüent de resultats a un o més motors de cerca, conscients que els usuaris finals sempre persegueixen, d'una forma o altra, la idea (en part utòpica) de la interfície de consulta universal. Naturalment, sindicació de continguts implica també un model de negoci. Implica que els motors de cerca com Google estiguin disposats a retribuir als productors de les bases de dades, o bé que, a partir d'un moment donat, una part dels resultats oferts pel sistema sigui d'accés lliure i una altra sigui d'accés condicionat al pagament d'una certa quantitat o a la condició de ser abonat o subscriptor.

Això és el que fa Scirus. Quan un usuari llança una cerca a Scirus pot trobar tres tipus de resultats: (1) documents d'accés totalment lliure, per exemple, un estudi publicat com una pàgina web en un servidor web convencional i d'accés lliure; (2) documents als quals té accés perquè la seva institució posseeix una subscripció a la publicació corresponent, per exemple un article d'una revista subscrita per la biblioteca de la seva institució; i (3) documents als quals té accés mitjançant pagament amb targeta de crèdit.

El Web semàntic

Definicions

Vegem primer la definició oficial de Web semàntic (*semantic web*), segons el W3 Consortium (l'organisme promotor de la idea):

"The Semantic Web is the representation of data on the World Wide Web. It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners. It is based on the Resource Description Framework (RDF), which integrates a variety of applications using XML for syntax and URIs for naming."

Dues coses sobre la definició anterior: en primer lloc, com es pot observar, no diu absolutament res. Què significa que alguna cosa és "la representació de dades en la *World Wide Web*"? Res. La resta de la suposada definició és pitjor. Abandona clarament l'intent de dir el que és el Web semàntic (veient l'antecedent, tal vegada sigui el millor) i es limita a assenyalar, entre d'altres coses summament informatives, "que integra una varietat d'aplicacions"(!).

La segona cosa que cal assenyalar és que el Web semàntic no existeix. No sabem si el Web semàntic serà realitat algun dia, però ara per ara, ni existeix ni se l'espera (almenys de manera imminent). Tot i així, s'ha de reconèixer en aquest concepte una autèntica idea-força, en el sentit que és una idea que ja ha estat capaç de mobilitzar moltes energies (i moltes il·lusions) i que, sens dubte, no deixarà de presentar resultats durant els propers anys, perquè segurament seguirà mobilitzant energies.

És una idea, per dir-ho d'alguna manera, semblant als viatges que tenen sentit per si mateixos, independentment de la destinació prevista. Diuen els experts en narrativa que tota autèntica aventura és en realitat un viatge en el qual, al final, el protagonista ha patit alguna transformació (se suposa que per a bé). El Web semàntic pot veure's com un viatge que inicia ara la *World Wide Web* i tal vegada no arribi mai (del tot) a la seva destinació, però que, mentre, la transformarà profundament.

Si haguéssim de proposar una definició de Web semàntic, nosaltres començaríem amb aquesta:

Definició: El Web Semàntic és un conjunt d'iniciatives, tecnològiques en la seva major part, destinades a crear una futura *World Wide Web* en la qual els ordinadors puguin processar la informació, és a dir, representar-la, trobar-la, gestionar-la, com si els ordinadors tinguessin intel·ligència

A continuació, intentarem presentar una aproximació a la idea del Web semàntic; per fer-ho, ens hem basat en un treball previ (Codina, 2003), però, sobretot, en la informació que sobre el Web semàntic pot trobar-se en el ja esmentat organisme promotor de la idea, el W3 Consortium (www.w3.org/2001/sw/), i en un famós i citadíssim article publicat a *Scientific American* (Berners-Llig, 2001). Hem consultat també altres autors que s'indiquen a la bibliografia.

Estat actual

Si el Web semàntic no existeix, què és en aquests moments? De moment, és el nom d'una aspiració; el nom d'un objectiu molt ambiciós que, de complir-se, canviaria de forma radical el Web tal com el coneixem avui. En què consisteix aquesta aspiració? Ni més ni menys es tracta d'aconseguir que les pàgines que formen el Web deixin de ser simples cadenes de caràcters per als ordinadors i

es converteixin en textos amb sentit, és a dir, text proveït de semàntica, tal com, de fet, ho és per als éssers humans.

Perquè un objectiu com aquest? Tal com es codifiquen les pàgines web actuals, principalment mitjançant el llenguatge HMTL, tenen molt poc sentit per a les màquines. En efecte, si veiem el codi font d'una pàgina web actual, trobem, per exemple, un tros de codi com aquest:

…Superar la bretxa digital…

Quan l'ordinador l'interpreti, a través del programa navegador, apareixerà com un text en negreta i cursiva com aquest:

…***Superar la brecha digital***…

Amb això gairebé s'acaba tot el que és capaç de fer un ordinador amb les pàgines HTML. Com saben els informàtics i documentalistes, una altra cosa que poden fer els ordinadors és construir índexs amb les paraules que apareixen en les pàgines web. Després, quan algú envia una pregunta a un motor de cerca, el que fa aquest últim és comparar les paraules de la pregunta amb les paraules del seu índex. Per exemple, suposem que el responsable d'un programa de govern sobre el problema de la bretxa digital decideix indagar a Internet per veure si troba estudis o informes sobre la bretxa digital.

Suposem que accedeix a Google i entra la següent pregunta: "bretxa digital". El que farà Google és comparar les paraules de la seva pregunta amb les paraules del seu índex. Si troba un document que tingui la "bretxa digital", el retornarà com a resposta. Això és gairebé tot el que poden fer els ordinadors que tingui a veure amb processament d'informació en pàgines web.

Amb aquestes limitacions, la cerca a Internet, com tothom sap, està plena de frustracions. Si algú cerca per "cavalls", no trobarà res que tracti sobre "egües". Si algú cerca sobre com evitar la guerra, no trobarà un document sobre com aconseguir la pau, etc. El Web semàntic vol solucionar això. Això sona a intel·ligència artificial. Per tant, encara que no vulguin anomenar-ho així, amb el Web semàntic s'està buscant el mateix objectiu, és a dir, que els ordinadors entenguin que un document sobre "egües" pot ser molt rellevant per a una necessitat d'informació sobre "cavalls", i que la semàntica de la pregunta "és possible evitar la guerra?" és la mateixa que la de la pregunta "és possible aconseguir la pau?".

A més, s'espera que els ordinadors puguin desenvolupar tasques de gestió que demanin interpretar informació i prendre decisions adaptant-les al context. Es tracta, ni més ni menys, d'un objectiu que la informàtica ha denominat fins ara intel·ligència artificial.

Infraestructura

Els mitjans amb els quals se suposa que s'aconseguirà el Web semàntic són els següents: primer, un nou llenguatge de codificació de pàgines, un nou llenguatge de marcatge. Aquest llenguatge, com és sabut, es denomina XML. Amb XML es poden dissenyar llenguatges de marcatge molt estructurats i molt explícits en els quals, en lloc d'etiquetes com <b> i <i>, hi haurà etiquetes com <títol>, <subtítol>, <capítol>, <subcapítol>, <autor>, <institució>, <ciutat>, etc.

Com que faran falta etiquetes específiques per a cada tipus d'informació –per exemple, les pàgines web de les companyies aèries necessitaran etiquetes com <vol>, <hora de sortida>, <destinació>, etc.– s'ha creat una especificació, una espècie de metallenguatge, XML, que permet definir llenguatges específics, és a dir, conjunts d'etiquetes específics per a cada necessitat d'informació. Per exemple, els editors de diaris disposen ja del seu propi conjunt d'etiquetes, així com els matemàtics per a expressar equacions, etc.

El segon element amb el qual es compta són les metadades. Com saben molt bé els documentalistes, les metadades són informació sobre la informació i són, en realitat, una antiga fórmula. Els catàlegs de les biblioteques són metadades. La venerable norma ISBD és una norma sobre metadades, els descriptors assignats a un document són metadades, els tesaurus i les classificacions són el que ara en l'argot de les metadades es denominen també *schemes*, etc.

La qüestió és que les pàgines web ja tenen metadades. Almenys, solen tenir la metadada títol, en forma d'etiqueta <*title*> en una zona de les pàgines web invisible per a les persones, però visible per als ordinadors. A més, algunes pàgines, molt poques, solen tenir altres metadades, com <*keyword*>, <*description*>, etc.

Com ja se sap, existeix una ambiciosa norma d'abast internacional, *Dublin Core*, que proporciona una llista unificada i normalitzada de fins a quinze metadades del tenor dels ja comentats, perquè els editors i autors que ho desitgin els incloguin en les seves pàgines web. La idea és simple: si les pàgines web tinguessin metadades del tipus <títol>, <autor>, <tema>, <lloc de publicació>, etc., els usuaris podríem fer preguntes molt més precises als motors de cerca. Podríem, per exemple, fer peticions d'informació d'aquest tipus: "busca'm documents publicats a tal lloc i que tractin d'aquest i aquest tema, sota aquest punt de vista".

Però les metadades actuals no tenen ni semàntica ni sintaxi ni estan unificades sota una norma comuna que agrupi la diversitat de plataformes de metadades existents.

Per dotar-les d'aquestes tres coses, s'han desenvolupat altres normes. La més important és la RDF (*Resource Description Framework*). Aquesta norma especifica una gramàtica lògica perquè els autors de pàgines web puguin descriure les propietats semàntiques dels documents en una notació estàndard i comuna per a qualsevol tipus de metadades. Es tracta d'una notació basada en nocions fonamentals. Bàsicament: hi ha objectes, com ara pàgines web, i els objectes tenen propietats, com un responsable intel·lectual, una data de publicació o un contingut expressat en paraules clau, etc. Així mateix, hi ha

relacions entre els objectes, com ara una pàgina web que forma part d'una sèrie o és una versió en altra llengua d'altra pàgina web, etc.

Per a descriure el contingut semàntic i altres propietats d'una pàgina web, es pot utilitzar la norma RDF mitjançant el procediment d'etiquetatge XML per expressar els temes d'un document, entre d'altres coses.

En síntesi, la gran esperança del Web semàntic es basa, almenys, en tres coses: XML per fer els documents més explícits; metadades (expressades també en XML) per fer els documents més fàcils de representar, indexar i buscar i, finalment –es desprèn de l'anterior, encara que sol obviar-se–, una nova generació de programari (*software*) –programes i mètodes de representació del coneixement– que sàpiga explotar les dues coses anteriors.

La representació del coneixement necessitarà, alhora, procediments normalitzats, ja sigui per representar coneixement complex o de sentit comú. Aquestes representacions solen denominar-se ontologies, un camp interdisciplinari on solen confluir diverses disciplines cognitives, des de la intel·ligència artificial fins a la lingüística.

Ara bé, en l'esquema del Web semàntic se suposa que les metadades les posen principalment els propis autors dels documents. Quin és el problema? En primer lloc, els autors no solen estar entrenats per posar metadades i es necessita molta formació per saber triar bones paraules clau.

En segon lloc, els autors –no tots, ni de bon tros– menteixen. Així de simple. Volen que les seves pàgines web quedin molt alt en els cercadors, de manera que col·loquen trenta vegades la mateixa paraula, amb petites variants, perquè quedin a la part de dalt dels rànquings dels motors de cerca per als temes que a ells els interessa, encara que la seva pàgina no tingui en realitat molt (o gens) a veure amb aquest tema.

En tercer lloc, les persones ens equivoquem, i els autors de les pàgines web s'equivoquen: s'obliden de posar metadades, les posen malament, les posen en unes pàgines sí i en unes altres no, s'equivoquen en l'ortografia, etc.

Conclusió: gairebé cap motor de cerca es fia de les metadades per generar els resultats dels seus rànquings.

Possibilitats reals a curt i mig termini

El lector ja haurà deduït que, almenys segons l'opinió de qui escriu això, les possibilitats a curt i mig termini del Web semàntic són reduïdes. Efectivament. Una cosa és que es tracti d'un objectiu que val la pena perseguir i l'altra que es tracti d'un objectiu factible. Permeteu-me un exemple molt significatiu: sens dubte és un bon objectiu (almenys, molts ho creiem així) acabar amb la pobresa al món. És un exemple d'una fita lloable, amb la qual tots hauríem de comprometre'ns. Però que sigui un objectiu magnífic i molt desitjable, no ho converteix automàticament en assolible; almenys no en la seva totalitat i no a mig o a curt termini. S'ha d'abandonar per això? Ni de bon tros. Tot el contrari. Cal perseguir-la amb afany, perquè és l'única forma d'aconseguir progressos en aquest terreny, encara que siguin parcials.

El problema amb el Web semàntic, tal com la presenten alguns dels seus defensors (sobretot el W3 Consortium, que sembla haver-se especialitzat a llançar confusió sobre tots els seus projectes recents) és la immensa quantitat d'ingenuïtat o d'ignorància que exhibeix. En comparació, els programes contra la pobresa i a favor dels drets humans són obres mestres de pragmatisme (i saviesa).

Seguim, per exemple, amb les metadades: si gairebé ningú utilitza metadades ara, per quina raó, de sobte, tot el món posarà metadades en les seves pàgines? A més, si els autors de pàgines web han demostrat la seva incapacitat per utilitzar una norma relativament simple com era la primera versió de *Dublin Core*, per què ho faran ara que ha dut la seva complexitat al límit del que és impracticable?

Finalment, respecte a les ontologies i la seva explotació mitjançant motors d'inferència o sistemes experts, si la intel·ligència artificial suma ja diverses dècades de fracassos, almenys en relació a la hipòtesi forta, és a dir en relació al seu objectiu declarat a so de bombo i platerets d'assolir que els ordinadors pensin, per què tindrà èxit ara?

Per tant, les possibilitats que el Web semàntic sigui una realitat tal com la presenta el W3 Consortium, sense que es produeixi abans, almenys, un canvi de paradigma en les ciències de la computació, són ridícules. A més, necessitarem en paral·lel canvis no menys importants en altres àrees, incloent, per descomptat, les ciències de la documentació.

Tanmateix, no ens enganyem, l'objectiu del Web semàntic és magnífic, produirà importants avenços en alguns o en tots els terrenys relacionats amb la representació i l'accés al coneixement i, al meu entendre, des de les ciències de la documentació, hauria d'obtenir tot el nostre suport.

Bolcadors, mapadors i altres eines de localització d'informació

Les eines de cerca de Segona Generació són programes client que automatitzen processos de localització, recerca i recuperació d'informació. Classificació:

- Bolcadors
- Multicercadors
- Traçadors
- Indexadors
- Mapadors de ports
- Continguts de la Infranet: catàlegs de biblioteques, bases de dades bibliogràfiques, obres de referència, estadístiques i bases de dades numèriques, o Bases de dades Textuals. Els agents de la Infranet són clients Z39.50, amb mecanismes per a la realització automàtica de cerques de forma simultània i que sol permetre el bolcat dels registres. Entre els directoris més interessants, destaquen:

- Directori de recursos Z39.50, a nivell internacional. http://www.ilrt.bris.ac.uk/discovery/z3950/resources/
- Directorio espanyol de recursos Z39.50. http://www.absysnet.com/recursos/recz3950.html. Mereix un especial esment Bookwhere, aplicació de cerca, recuperació i exportació de la informació que utilitza el protocol Z39.50 d'Internet, i que té com a objectiu facilitar l'accés a registres bibliogràfics i a text complet via Internet. (Versió Demo a: http://www.web-clarity.com/products/overviewbookwhere.html).
- Continguts del Web invisible: pàgines òrfenes (sense connexió hipertextual); pàgines no textuals (com fitxers multimèdia i executables); pàgines amb accés mitjançant passarel·les (com pàgines amb paraula clau d'accés, ja siguin gratuïtes o de pagament; dipòsits de documents; revistes electròniques, etc.) o pàgines dinàmiques. Algunes adreces que permeten accés als continguts d'Internet invisible són: www.internetinvisible.com i www.completeplanet.com

Conclusions

En el futur dels sistemes d'informació hi ha una llarga llista d'innovacions a les quals val la pena parar atenció. Assenyalarem les que són més importants, segons la nostra opinió, pel seu impacte en les Ciències de la Documentació:

1. **Internet invisible**. S'ha produït un gran avenç en la varietat de formats que poden indexar els motors de recerca. D'altra banda, és possible que motors de cerca com Scirus siguin només un exemple de la classe de sistemes d'accés a la informació que podem esperar en el futur. Tot i així, hi ha diversos fronts en què hauríem de començar a col·locar les nostres energies i esforços. D'una banda, els documents no HTML són potencials enemics de la hipertextualidad. Hauríem de considerar si els avenços per una banda, no són retrocessos per una altra. En aquest cas, hauríem de considerar què fer o, almenys, considerar què fer en el camp de la investigació i les polítiques d'informació. Segur que tenim un ampli i bonic programa d'investigació per aquest costat. D'altra banda, les interfícies de consulta dels motors de cerca estan a anys llum de les possibilitats reals i del *know-how* sobre el tema. Un altre punt sobre el qual cal pensar o, millor encara, actuar.

2. **Web semàntic**. Encara que sigui amb mentalitat d'ONG, què podem fer a favor del Web semàntic si creiem en els seus beneficis a escala social encara que, ara com ara, aporti escassos beneficis individuals? Els organismes vinculats al món de la promoció del coneixement i la ciència i el patrimoni cultural (universitats, arxius, biblioteques, centres d'investigació, museus, etc.) s'haurien d'interessar-se pel Web semàntic. Per tant, a curt i mig termini, les organitzacions vinculades amb el món de la ciència, la cultura, el patrimoni, l'educació, etc., haurien de sentir-se obligades a: (1) interessar-se almenys per

coses tan aparentment innocents com el llenguatge XHTML juntament amb els fulls d'estil (CSS) i (2) estudiar polítiques de metadades en relació a totes les seves publicacions digitals.

3. **Què ens ensenya el Web semàntic?** Al meu entendre, ens ensenya el que, en realitat, ja sabíem: si agafes un conjunt de dades i les etiquetes sistemàticament i exhaustivament, tens el més semblat a la intel·ligència. Si les bases de dades exhibeixen un notable grau d'intel·ligència en comparació al Web és perquè en una base de dades tots les dades estan "etiquetades", és a dir, formen part dels valors d'un camp. Cada camp, al seu torn, té uns atributs ben definits: és un camp de text o és un camp numèric, o lògic, etc. Finalment, tots les dades en una base de dades estan sistematitzades: cada registre respon a la mateixa estructura, així que la posició (la sintaxi) genera sentit (semàntica). Així que, el que és (genialment) nou en el Web semàntic és la idea de convertir tota el Web en la més gegantesca base de dades que la humanitat hagués somiat mai.

Iniciatives de patrimoni digital

Projecte de la Carta de la UNESCO per a la Preservació del Patrimoni Digital.

Diu la UNESCO en el seu document "DIRECTRIUS PER A LA PRESERVACIÓ DEL PATRIMONI DIGITAL":

"Gran part de la ingent quantitat d'informació que es ***produeix*** en el món és d'origen digital i existeix en una gran varietat de formats: text, bases de dades, enregistraments sonors, pel·lícules, imatges. Per a les institucions culturals que tenen al seu càrrec la ***recol·lecció*** i la preservació del patrimoni cultural, definir quins elements han de conservar-se per a les generacions futures i com procedir en la seva selecció i conservació, s'està tornant un problema urgent. L'enorme tresor d'informació digital produïda avui dia en pràcticament totes les àrees de les activitats humanes i concebuda per a ser consultada amb computadores, podria perdre's si no s'elaboren tècniques i ***polítiques específiques per a seva*** conservació**.**"

Podeu llegir el text complet en:

http://unesdoc.unesco.org/images/0013/001300/130071s.pdf

➔ F. Gestió del coneixement i eines col·laboratives

El coneixement es dóna quan una persona coneix o sap alguna cosa d'algun tema, però no hi ha dubte que els avenços de la ciència han contribuït amb la

formació del coneixement. Com tots sabem, el coneixement es forma a través de l'experiència que adquirim amb el pas del temps, ja que el coneixement es forma a través de la informació que tenim a l'abast de les nostres mans gràcies a Internet, que ens ha ajudat molt en el camp de la ciència i tecnologia, i en la nostra vida diària.

Tots sabem que la Wikipedia és una enciclopèdia que està a l'abast de tots a Internet, ja que és gratuïta i podem editar-la i redactar algun article sobre un tema del qual tinguem coneixements. El seu principal avantatge és que ens proporcionar informació de manera més ràpida i així podem adquirir nous coneixements. Aquesta eina permet que cada alumne, des d'allà on es trobi, pugui investigar, redactar i publicar la informació que posseeixi i, al mateix temps, llegir la informació que aporten els seus companys. Finalment, una posterior edició dels continguts permetria crear una definició col·lectiva i probablement molt més rica (sota el principi d'intel·ligència col·lectiva) que la que cada estudiant ha redactat individualment.

WordPress comprèn els blocs, que també són fonts d'informació, ja que podem pujar-hi informació. Això s'a convertit en una bitàcola per a estudiantes i professors en el procés educatiu, ja que és un espai per escriure preguntes, publicar treballs o registrar enllaços cap a recursos rellevants. Actualment, existeixen nombroses comunitats de bloc educatives on s'intercanvia informació i coneixement entre professors i alumnes. Aquest tipus de pàgina web d'estructura cronològica s'ha convertit en el sistema de gestió de continguts més popular del Web 2.0 i un dels favorits de molts professors. Aquesta eina és molt important, ja que podem tenir accés a qualsevol tipus d'informació, sempre que la sapiguem utilitzar correctament.

You Tube: aquí trobem vídeos educatius i d'entreteniment.

Flickr: per compartir fotografies o imatges.

Col·laboratoris: aquest tipus de plataformes s'utilitzen com a repositoris per a l'educació, ja que permeten compartir objectes d'aprenentatge que després poden exportar-se a altres plataformes. Són també espais de cooperació per al desenvolupament d'investigacions. Els col·laboratoris simplifiquen de manera notable l'accés i intercanvi d'informació entre professors, acadèmics i estudiants, com si fos una biblioteca o un laboratori de lliure accés. Aquí es poden compartir documents científics, projectes, reportis, conferències, papers, classes, tasques, estudis, bases de dades, etc.

Avantatges

- Un dels avantatges és que es trenca el model de programari tancat amb drets d'ús i sota el principi de la obsolescència planificada, per passar a l'ús del programari gratuït o lliure
- Els recursos en línia del web 2.0 optimitzen la gestió de la informació, que es converteix en instruments que afavoreixen la conformació de xarxes d'innovació i generació de coneixements basats en cooperació i reciprocitat.

- El desenvolupament d'habilitats en els educands estimula el seu interès per generar i compartir recursos multimèdia de qualitat.

Desavantatges

- Un desavantatge, que pot arribar a ser el més important, és que nosaltres mateixos ens creem una dependència a l'ús d'Internet. A més, existeix una polèmica al voltant de la rellevància i pertinença del terme Web 2.0, fins al punt de ser qüestionat per molts actors del propi entorn, que consideren que Web 2.0 és la denominació més apropiada per descriure el nou tipus d'aplicacions web dominants i la fase actual en la qual es troba la xarxa creada per Berners-Lee. El problema és que les notícies més llegides o votades no són les més importants, mentre que les que realment ens serviran d'ajuda no són tingudes en compte per la major part d'usuaris.
- La propietat intel·lectual és un tema important, però alguns usuaris de Wikipedia i blocs no saben a què es refereix aquest terme i simplement busquen i descarreguen la informació que necessiten però no citen el nom de l'autor, i així es perd el nom o la identitat de l'autor de la informació que agafem i això és perjudicial, ja que, si no citem l'autor original o la font de la informació que estem publicant o utilitzant, ens poden sotmetre a sancions econòmiques o privar-nos de la llibertat.

Crowdsourcing

És un concepte que sorgeix d'aprofitar l'arquitectura social del Web 2.0, els nivells creixents de participació mediatitzada i el poder de la intel·ligència col·lectiva, la suma dels quals s'ha convertit en una font d'idees i desenvolupaments per al sector empresarial i fins i tot per al camp de l'experimentació científica. A més, es refereix a una altra manera d'emprar mà d'obra barata, gràcies a la popularitat i ubiqüitat d'Internet, en la qual persones no especialitzats resolen problemes per a tota classe de companyies que utilitzen el potencial dels milions de cervells de la multitud que es connecta a través de la Xarxa. Crowd és el terme en anglès de multitud i sourcing es refereix a l'obtenció de matèria primera (on source és el terme en anglès de font, en aquest cas d'un projecte).

Tipus de treball massiu

La wiki és la crowdsourcing més coneguda:

- Worthidea és una plataforma multicultural i multilingüe d'idees que servei com a punt de trobada entre les empreses i els usuaris.
- Procter and Gamble té més de 9.000 científics i investigadors treballant a la corporación R&D i encara tenen molts problemes que no poden solucionar. Ells ara publiquen els seus terribles mals de cap en un lloc anomenat InnoCentive, oferint grans sumes de diners a més de 90.000

"solucionadors", que tenen la seva xarxa de científics de suport. P&G també treballa amb NineSigma, YourEncore i Yet2 .

- ➔ Amazon realitza una cosa similar amb projectes de sotfware a gran escala.
- ➔ iStockphoto és un lloc amb més de 22.000 fotògrafs amateur, que pugen i distribueixen estocs fotogràfics. Com que no tenen els alts costos d'un equip professional como Getty Images, és possible comprar imatges por un baix cost. Aquesta empresa va ser comprada per Getty Images.
- ➔ CambriaHouse és un headquarter de iStockphoto a Calgary, Canadá, i es descriu com: codi obert amb diner. És una incubadora que descobreix i comercialitza programari i idees a través del crowdsourcing. Els contribuents guanyen regalies i comparteixen els guanys del producte.
- ➔ Portucuenta és un portal de programadors freelance en espanyol, que permet desenvolupar projectes de qualsevol escala, ja que compta amb un pool de programadors disponibles als quals se'ls pot assignar diferents tasques per integrar-les després.

Open Innovation

Open Innovation és un terme promogut per Henry Chesbrough, professor i director executiu del Centre d'Innovació Oberta de Berkeley.

La idea central darrere d'*Open Innovation* és que en un món de coneixements d'àmplia distribució, les empreses no poden permetre's el luxe de dependre només de la seva pròpia investigació, sinó que han de comprar llicències de processos o invencions (patents) d'altres empreses. A més, les invencions internes que no s'utilitzen s'haurien de portar fora, per exemple a través de la concessió de llicències, les empreses mixtes, les *spin-offs*, etc.

Avantatges

- ➔ *Open Innovation* parteix de la premissa que la informació i el coneixement són abundants i estan àmpliament distribuïts.
- ➔ En els models anteriors, el lloc de la innovació era l'empresa. *Open Innovation* entén que els actors interns i externs tenen un paper similar.
- ➔ El paper central del model de negoci en tot el procés. En estructures anteriors el model de negoci tenia un paper secundari en el procés d'innovació. A *Open Innovation* el model de negoci té un paper dual: a) la selecció de productes i serveis pels quals apostar, i b) la recerca i la creació activa de models que permetin comercialitzar aquelles idees que no encaixen en el model de negoci actual.
- ➔ *Open Innovation* considera projectes encara que no encaixin en el model de negoci. Aquests projectes poden ser rellevants ja sigui perquè es dirigeixen al propi mercat o a mercats potencials on podem capturar valor.

- *Open Innovation* entén la innovació com un procés global, d'aquesta manera les unitats de negoci no només competeixen internament sinó també amb l'exterior.
- Un paper proactiu de la gestió de la IP (sigla anglesa de propietat intel·lectual) a través de llicències, llicències creuades o fins i tot donació de patents.
- Un conjunt de mètriques noves en l'avaluació del procés d'innovació, d'acord amb el canvi de *locus* i la comprensió global del procés que proporciona el nou model (activitats d'innovació fora de l'empresa, nombre de *partnerships*, nombre d'*spin-offs*, etc.).

Tipologia d'eines col·laboratives

Aquesta classificació es fa prenent com a referència la classificació realitzada per McGreal, Gram i Marks, que van tenir en compte treballs realitzats amb professionals de l'educació:

- Eines per a la gestió i administració acadèmica: es gestionen assumptes com la gestió de la matrícula i inscripció dels alumnes en els cursos i proporcionar informació acadèmica com horaris, dates d'exàmens, notes, plans d'estudis, expedició de certificats, concretar reunions, tutories, etc. A la Universitat de Antioquia, per exemple, això s'assoleix a través de MARES.
- Eines per a la creació de materials d'aprenentatge multimèdia: aquí es poden trobar aquells programes que són utilitzats per a la creació dels continguts d'aprenentatge: els editors de pàgines web com HTML o els que faciliten la creació d'exercicis d'autoavaluació, simulacions, o pràctiques, com la creació de wikis o la realització de mapes conceptuals en línia.
- Eines per a la comunicació i el treball col·laboratiu: en aquest grup es poden trobar aquelles que faciliten la comunicació a través d'un ordinador entre alumne-professor, com el correu electrònic, els xats, les conferències electròniques, les àudioconferències, les videoconferències, la pissarra compartida, aplicacions compartides o documents compartits.
- Eines integrades per a la creació i distribució de cursos a través del WWW. Desenvolupades específicament per a propòsits educatius. En aquest grup es poden trobar totes aquelles plataformes que ofereixen cursos en línia, per exemple.

Altres tipus:

- E-mail: serveix per rebre i enviar missatges de manera immediata.

- Backpack: és una aplicació per organitzar el material per elaborar els projectes compartits. La informació sobre aquesta aplicació es pot trobar a http://www.backpackit.com/.
- Vyew: serveix per crear una sala de Xat amb un número de sala que de quatre dígits com a màxim. Aquest número es comparteix amb les persones amb les quals es vol realitzar el projecte. En aquesta eina es poden crear icones d'organització i emmagatzematge per a la informació que es vagi produint. Si voleu un exemple, podeu visitar http://vyew.com/.
- G talk: aquesta eina, creada per Google, serveix per enviar i rebre missatges instantanis entre dues o més persones. Només cal tenir un compte a Gmail. Per observar més beneficis o aplicacions d'aquesta eina, es pot consultar http://www.google.cat/talk/.
- Weblocs o bitàcoles: es tracta d'una pàgina web amb apunts datats en ordre cronològic invers, perquè l'usuari pugui trobar en primer lloc les últimes informacions.
- Wikis: un wiki és un exemple clar i precís del que és treball en grup, ja que és una eina creada i mantinguda per diversos autors, fet que la diferencia dels weblocs, que només poden ser modificats per l'autor original.
- Xarxes socials: les xarxes socials també són conegudes com *software* socials.

En l'actualitat, totes aquestes eines s'estan incorporant en les unitats d'informació d'una manera vertiginosa, però cada institució va al seu ritme depenent dels recursos econòmics i tecnològics que té. Ara bé, independentment que es compti o no amb els recursos suficients perquè les unitats d'informació treballin amb les eines col·laboratives, és necessari que es comenci a interactuar amb aquestes, ja que és una manera molt pràctica perquè la unitat d'informació es retroalimenti amb les inquietuds i encerts dels usuaris en línia. A més, les eines col·laboratives permeten a les unitats d'informació emprendre projectes amb d'altres i així ampliar més les seves perspectives, fet que redundarà en beneficis per als seus usuaris tant reals com potencials. Personalment recomanaria els *wikis*, els weblocs i els *backpacks*, almenys mentre es comença a aprofundir en altres eines.