Eines de cerca

Ana María Checa Rubio
Pere Masip Masip

P06/09046/00263

UOC
www.uoc.edu

Índex

1. La cerca d'informació a Internet

La cerca i recuperació de la informació a Internet presenta unes característiques diferents respecte a la que es realitza en els serveis d'informació tradicionals, com les biblioteques i els centres de documentació.

Aquest diferent funcionament s'explica per les particularitats dels recursos d'informació a Internet, que (Burnet i altres, 1999):

- no són fixos i estables, com els documents impresos o enregistrats en algun tipus de suport físic.
- no estan seleccionats i organitzats.
- no estan organitzats d'una forma centralitzada que en faciliti l'accés, no són catalogats, ni indexats.

A les quals, hi caldria afegir:

- **Volum**. La Xarxa posa a la nostra disposició una gran quantitat d'informació, que creix constantment.

  El volum d'informació a Internet

  Segons un estudi realitzat per investigadors de la Universitat de Berkeley (EUA), es calcula que el volum d'informació a Internet és de 532.897 terabytes. Un terabyte equival a 1.000.000.000.000 bytes.

  Font: P. Lyman; H. R. Varian (2003). *How much information* (document electrònic, consulta: 21/03/05)

- **Varietat**. La informació (textual, sonora, audiovisual...) ens arriba per diversos mitjans, com el correu electrònic, els xats, les llistes de distribució, el web..., i en múltiples formats: HTML, PDF, DOC, GIF, JPEG, WAV, MP3...
- **Volatilitat**. Aquesta informació està permanentment disponible, alterable i actualitzable. És freqüent observar com pàgines web que estan plenament operatives i en funcionament, en poc temps hagin desaparegut o modificat completament o parcialment el seu contingut. La vida mitjana d'un web és de 2,9 anys (Koehler, 1990).

Les especials característiques que ofereix la Xarxa van provocar que, pràcticament des dels seu naixement, sorgissin enginys encaminats a facilitar la localització i recuperació de la informació. Tanmateix, no va ser fins a la popularització del web que es va generalitzar l'ús de les eines de cerca i recuperació d'informació a Internet, conegudes genèricament com a *cercadors*.

2. Models de recuperació de la informació a Internet

La recuperació de la informació a Internet es fonamenta en dos paradigmes principals:

- El *browsing* o navegació, que se situa en la base dels directoris com ara *Yahoo*.

- La interrogació per paraules clau, que distingeix els motors de cerca, com, per exemple *Google*.

2.1. *Browsing* o navegació

Comporta elegir successivament, de manera no sistemàtica, per mitjà d'una determinada estructura d'informació existent, com un quadre de classificació o una estructura de categories, fins que es localitza el que es busca.

Chowdhury defineix la navegació com:

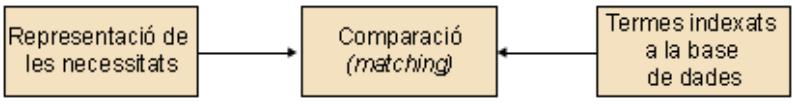
> "Una activitat de cerca interactiva en la qual la direcció de la cerca està determinada per l'usuari sobre la base d'una immediata retroalimentació sobre el que es navega."
>
> G. G. Chowdhury (1999). *Introduction to modern information retrieval* (pàg. 302). Londres: Library Association Publishing.

Com s'ha dit, la navegació és el tipus de recuperació de la informació que fonamenta els directoris d'Internet.

2.2. Interrogació

La interrogació consisteix a expressar una necessitat d'informació a una base de dades i com a resposta a la petició, el sistema ofereix els resultats que, en principi, haurien de satisfer millor aquelles necessitats.

Les necessitats d'informació es representen per mitjà d'un o més termes que conformen una expressió de cerca, i la resposta dels sistema seria fruit de l'aparellament (*matching*) dels termes que representen les necessitats amb els termes emmagatzemats a la base de dades.



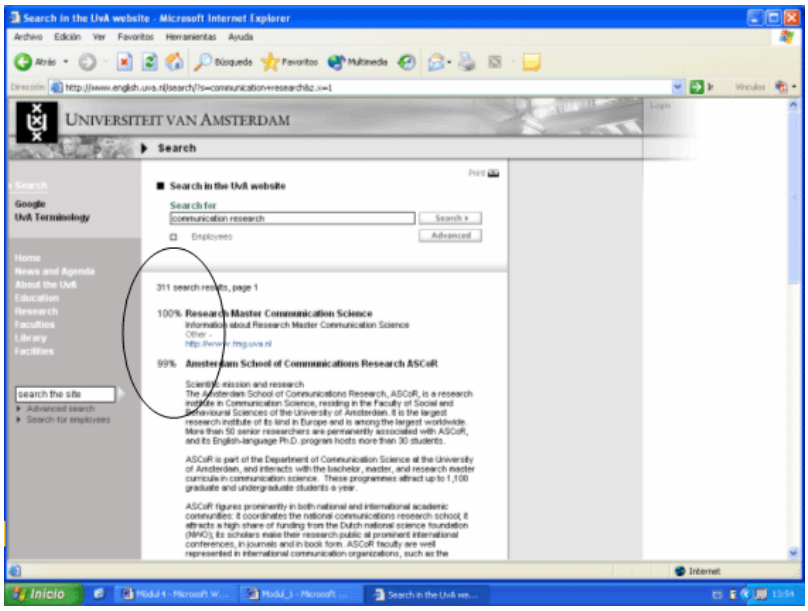
La interrogació de la informació es troba en la base dels motors de cerca com ara Google, però també en els catàlegs i les bases de dades. Mentre que en aquests darrers, els resultats de la interrogació se solen presentar ordenats alfabèticament o cronològicament (un sistema poc eficient tanmateix), en els motors de cerca els resultats s'ordenen segons la seva rellevància, de manera que els documents que el sistema considera més rellevants s'ofereixen més amunt que els que ho són menys.

Rellevància

El concepte de *rellevància* ocupa un paper protagonista en la recuperació de la informació a Internet. D'una manera senzilla, es pot definir rellevància com la relació entre la necessitat d'informació manifesta i la informació emmagatzemada. Un document serà més rellevant com millor pugui satisfer una necessitat informativa.

La rellevància té un grau; és a dir, no tots els documents són igualment rellevants: un document pot ser més o menys rellevant a una pregunta. Si s'accepta aquest supòsit, que diferents documents tenen diferents graus de rellevància, els sistemes de recuperació de la informació han d'intentar determinar-la de la forma més eficient possible, oferint els documents considerats més rellevants abans que els recollits com a menys rellevants.

Rellevància en un cercador



Cercador del web de la Universitat d'Amsterdam. Estimació de la rellevància dels documents.

Els motors de cerca, com s'ha comentat, són els exemples més característics de la recuperació de la informació a Internet basada en la interrogació. Tradicionalment, els càlculs de rellevància usats pels cercadors s'han basat en els elements següents:

- Freqüència absoluta d'aparició dels termes de cerca. És el sumatori del nombre de vegades que el terme o termes buscats apareixen en un document. Com més gran sigui la freqüència, més rellevant serà el document.

- Freqüència relativa d'aparició dels termes de cerca. És el producte de la freqüència absoluta pel nombre total de paraules d'un document. Com més gran sigui la freqüència, més rellevant serà el document.

- Ubicació. Situació del terme o termes cercats en el conjunt del document. Així, un document que contingui el terme buscat en el títol o en les metaetiquetes serà considerat més rellevant que un que el tingui entremig d'un paràgraf del text.

- Proximitat entre termes. En una cerca que contingui més d'una paraula clau, com més proximitat en el text dels diversos termes, més rellevància del document.

La major part de motors de cerca realitzen els càlculs de rellevància a partir de la combinació d'aquests quatre criteris principals.

El panorama dels motors de cerca va canviar substancialment gràcies a l'aparició de Google l'any 1998. El cercador creat per Larri Page i Sergey Brin, a més del criteris tradicionals per al càlcul de la rellevància, va afegir un criteri extern al document. En concret, el nombre i la qualitat d'enllaços que rep un document. Dit en altres paraules, va incorporar un criteri basat en la popularitat d'una pàgina web. Aquest sistema de mesura de la rellevància, Google l'anomena ***PageRank.***

Segons aquest criteri, com més enllaços externs rebi una pàgina web, més rellevant es considerarà.

Google, a més, no atorga el mateix valor a tots els enllaços. Rebre un enllaç d'una pàgina que, al seu torn, rebi més enllaços, serà més valorat a l'hora de realitzar el càlcul de rellevància que l'enllaç provinent d'una pàgina menys popular.

Alguns motors de cerca usen criteris addicionals per a calcular la rellevància. Alexa, per exemple, valora també el nombre de visites que rep una pàgina. Una pàgina amb més visites serà més rellevant que una amb menor nombre de consultes. Aquest criteri es coneix amb el nom de ***TrafficRank.***

Posicionament

Entenem per *posicionament* la capacitat de col·locar una pàgina web en una situació privilegiada dins dels resultats proporcionats per un motor de cerca. Per a aconseguir-ho, els administradors de webs poden optimitzar les pàgines web i usar els diversos procediments i tècniques que tenen al seu abast.

Per a saber-ne més: la revista *El Profesional de la información* (vol. 14, núm. 1 i 2 de 2005) va dedicar dos números monogràfics al tema del posicionament.

3. Els cercadors: tipologies

Seguint els paradigmes principals de recuperació de la informació a Internet, a grans trets, podem distingir tres tipus principals d'eines de localització d'informació a la Xarxa:

- Directoris o llistes jeràrquiques.
- Motors de cerca.
- Metacercadors.

Bases de dades

Alguns autors inclouen en aquesta categorització les bases de dades, en el nostre cas s'han exclòs, ja que és una tipologia ja existent abans de la irrupció de la Xarxa.

3.1. Directoris

Els directoris consisteixen en una recopilació de recursos web organitzats de forma lògica i seguint una classificació o estructura jeràrquica de categories, que va des de les més genèriques, fins a les més específiques.

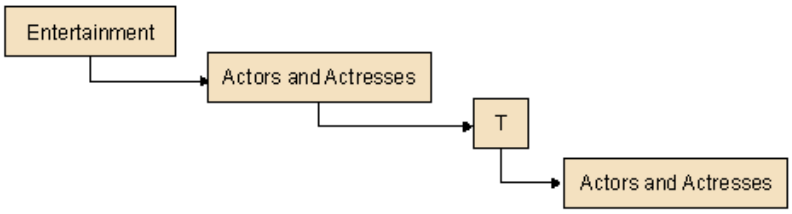
Categories de Yahoo

Així, per exemple, a Yahoo, probablement l'exemple de directori més paradigmàtic, els recursos web s'estructuren a partir de catorze categories principals, les quals, al seu torn, se subdivideixen progressivament en d'altres de més específiques.

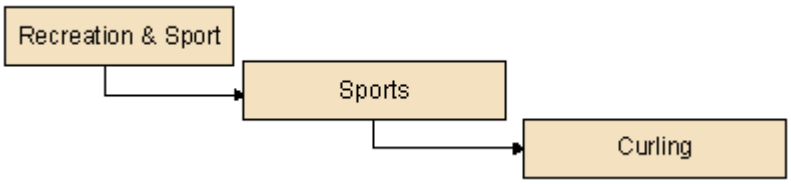
La cerca es realitza mitjançant nagevació o *browsing*. A partir de les categories principals es descendeix en l'estructura jeràrquica que caracteritza els directoris fins a categories més específiques, i fins a arribar a la categoria més precisa que aplega les webs sobre la informació que es busca.

Categories de Yahoo

Les pàgines web sobre Quentin Tarantino a Yahoo es troben en les categories següents:



I si es necessita informació sobre el Curling, caldrà navegar per les categories següents:



Les categories que constitueixen l'estructura jeràrquica dels directoris són, en la major part dels casos, categories arbitràries, creades *ad hoc* per cada cercador. Tanmateix, hi ha alguns directoris especialitzats que han adoptat classificacions bibliotecàries com la DDC o l'LCC per a organitzar els seus recursos.

Vegeu

Sobre classificacions bibliotecàries com la DDC o l'LCC per a organitzar els seus recursos, podeu anar al mòdul "La importància del llenguatge en la recuperació de la informació" d'aquest material didàctic.

La inclusió dels recursos en els directoris i la seva classificació en la categoria més adequada, la realitzen persones, ja siguin els mateixos creadors dels recursos que proposen la categoria més idònia o els especialistes dels directoris encarregats de fer-ho.

A causa del creixement del volum d'informació disponible al web i el sistema d'inclusió i classificació que caracteritza els directoris, aquests no tenen una voluntat de ser exhaustius; és a dir, no vol aplegar tota la informació disponible al web. En aquest propòsit, en teoria, sí que buscarien els motors de cerca, tot i que de forma quimèrica.

Directoris i motors de cerca

Els directoris són més precisos que els motors de cerca, però menys exhaustius.

Per la seva estructura, els índexs temàtics són més adequats:

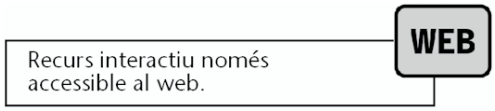
- Per a realitzar cerques de temàtiques generals.
- Per a realitzar cerques sobre temes que no es coneixen en profunditat.
- Per a localitzar els recursos considerats més importants sobre un tema.
- Quan no es té clara l'estratègia de cerca que es vulgui realitzar.

Els directoris, però, no estan exempts de problemes:

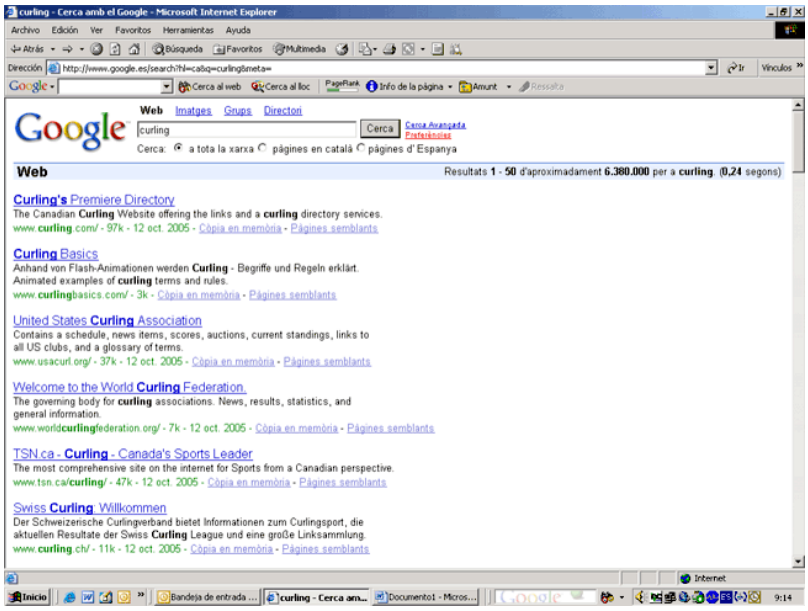
- Les categories són arbitràries, fet que pot generar problemes a l'hora de saber en quina categoria buscar.
- L'actualització no es fa de forma regular i és lenta.
- Els criteris d'avaluació i inclusió no sempre són públics, transparents i prou establerts.

Ús d'un directori o d'un motor de cerca?

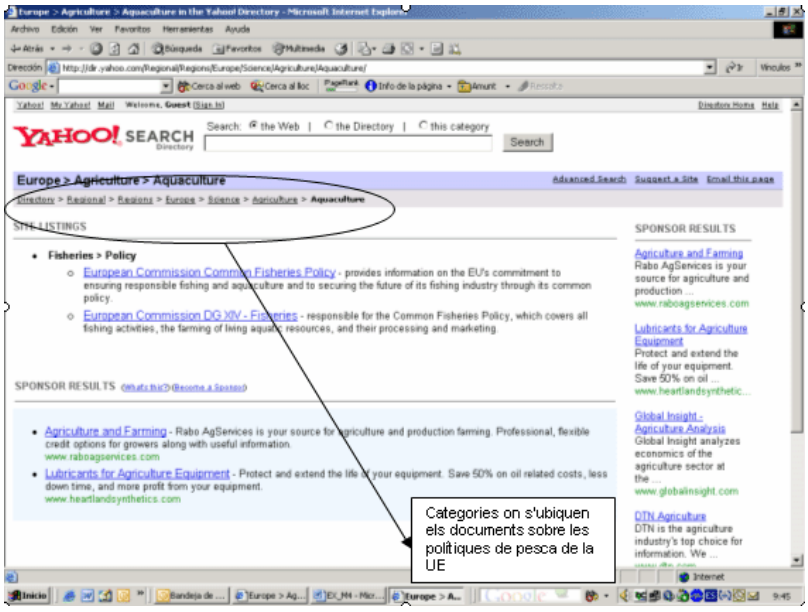
L'ús d'un directori és especialment indicat si es cerca informació general sobre un tema. En l'exemple següent es mostra els resultats obtinguts després de cercar a Yahoo informació sobre el *curling* (exemple anterior). Cal observar que els resultats són pocs, no arriben al centenar, i força generals (organismes vinculats a aquest esport, clubs...), però adequats per a una persona que vol una informació introductòria al tema.



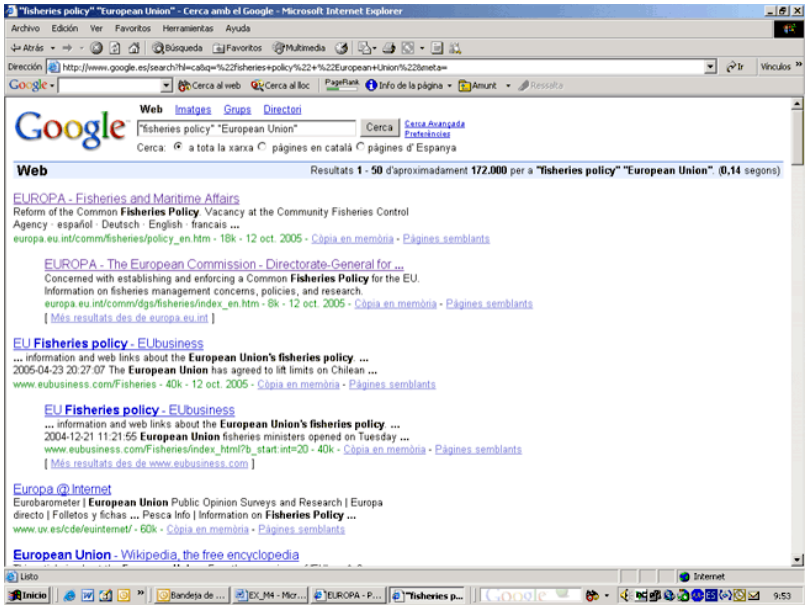
Si realitzem una cerca a Google per a localitzar informació sobre el *curling*, els resultats són més exhaustius i variats, però el gran volum de pàgines recuperades i la seva heterogeneïtat poden representar un problema per a un usuari desconeixedor del tema i que únicament necessita informació bàsica introductòria.



Si els directoris poden ser un recurs a considerar davant una necessitat d'informació de tipus general, presenten molts més inconvenients quan el tema de cerca és específic. Així, per exemple, si un usuari necessita informació sobre la *Política de pesca de la UE* ha de navegar per sis nivells de profunditat en l'estructura jeràrquica de Yahoo abans de trobar la informació que necessita. A més, a causa del caràcter arbitrari de les categories de la major part dels directoris, no sempre resulta fàcil saber en quina categoria buscar.



Davant una necessitat d'aquest tipus, la consulta d'un motor de cerca pot resultar més profitosa.



A causa del gran nombre d'enllaços que apleguen alguns directori, juntament amb la característica estructura jeràrquica, la major part dels directoris, actualment, ofereixen també motors interns de cerca que permeten buscar dins la base de dades usant paraules clau. Aquest sistema facilita l'accés a la informació, però pot desvirtuar el principal valor afegit que proporcionen els directoris.

Exemples de directoris

Exemples de directoris	
Yahoo	http://dir.yahoo.com
Open Directory	http://www.dmoz.org

Exemples de directoris	
Looksmart	http://www.looksmart.com

3.1.1. Biblioteques virtuals

Derivades del concepte de directori, aviat van proliferar eines de cerca que recopilaven recursos de forma selectiva per la seva qualitat i utilitat seguint criteris rigorosos i professionals. Aquests directoris sovint reben el nom de *biblioteques virtuals* (no s'han de confondre amb les biblioteques digitals), tot i que també apareixen en la literatura especialitzada com a directoris especialitzats, *subject gateways, etc*.

El nombre de recursos que aplega un directori d'aquest tipus és molt menor que el que pot aplegar qualsevol altre cercador, però gràcies al seu caràcter selectiu, l'ús de les biblioteques virtuals permet augmentar la possibilitat d'obtenir resultats rellevants.

A grans trets, les biblioteques virtuals:

- Són l'evolució dels grans directoris generalistes.
- Originalment eren impulsades per biblioteques universitàries o organismes sense finalitat de lucre.
- Els recursos se seleccionen selectivament atenent a criteris de qualitat i utilitat.
- Sovint els criteris de selecció són públics.
- Tenen una mida petita.

Exemples de biblioteques virtuals

Exemples de biblioteques virtuals	
Generalistes	
BUBL	http://bubl.ac.uk
The Argus Clearinghouse	http://www.clearinghouse.net
Librarians's Index to the Internet	http://lii.org
Internet Public Library	http://www.ipl.org
Especialitzades	
SOSIG (ciències socials)	http://www.sosig.ac.uk
MCS	http://www.aber.com.uk/media
Portal de la comunicació	http://www.portaldelacomunicacion.com

3.2. Motors de cerca

Els motors de cerca permeten la cerca d'informació al web a partir de paraules clau gràcies a l'existència d'una base de dades que indexa la informació que rep de programes informàtics anomenats genèricament robots o *spiders*.

Els motors de cerca estan constituïts per quatre elements principals:

- **Robot o *spider*** . Aplicació web que recorre constantment i de forma automàtica la Xarxa amb la finalitat de localitzar documents web. Un cop detecta un document, n'envia una còpia al programa d'indexació que utilitza el motor de cerca per a generar una representació del document i incorporar-lo a la base de dades. Els robots usen els enllaços hipertextuals inclosos en els documents web per a desplaçar-se fins a recursos nous.

- **Programa d'indexació**. La informació que els motors de cerca reben provinent dels robots no es pot incorporar directament a la base de dades, prèviament un programa d'indexació automàtica elabora un índex invers que permetrà la posterior recuperació dels documents a partir del seu contingut.
  Els índexs inversos estan formats majoritàriament per paraules clau i la seva ubicació en el text; ja hi ha, però, intents d'elaborar una indexació per conceptes, d'acord amb teories lingüístiques i a la intel·ligència artificial. Aquesta línia de treball es coneix com a ***web semàntic.***

- **Base de dades**. Està constituïda per l'índex invers creat pel programa d'indexació i l'URL de tots els documents indexats.

- **Interfície de l'usuari**. És l'element que permet la interacció entre l'usuari i la base de dades. Mitjançant la interfície, l'usuari realitza la consulta del recurs, introduint l'equació de cerca que serà interpretada pel llenguatge d'interrogació. Posteriorment, i després de comparar els termes que representen les necessitats informatives amb els termes emmagatzemats a la base de dades, el sistema retorna els documents que satisfan la demanda de l'usuari.

Habitualment cada motor de cerca sol tenir el seu propi llenguatge d'interrogació, tot i que es detecta una tendència vers l'estandardització. Malauradament, l'estandardització condueix a la simplificació. Així, per exemple, motors de cerca amb llenguatges d'interrogació potents, com Altavista, han deixat d'oferir la possibilitat d'usar operadors com els de proximitat o les funcions de discriminació per majúscules o accents.

En comparació dels llenguatges d'interrogació de les bases de dades tradicionals o dels catàlegs, ja siguin en entorns web o no, els llenguatges d'interrogació dels motors de cerca ofereixen moltes menys prestacions. Aquesta simplificació i reducció de les possibilitats de cerca, pot generar la falsa imatge que la cerca i recuperació de la informació a Internet són una qüestió senzilla.

Els **principals avantatges** que ofereixen els motors de cerca són:

- s'actualitzen freqüentment,
- permeten l'accés a una gran quantitat d'informació i
- faciliten un resultat més exhaustiu que un directori.

Amb tot, no estan exempts de **problemes**:

- És necessari el coneixement de la sintaxi d'interrogació de cada cercador.

- Poden generar desbordament cognitiu; és a dir, la quantitat d'informació que proporcionen és tan important que es pot fer difícil abordar-la i, evidentment, assimilar-la. Aquesta situació pot generar ansietat.

L'ús dels motors de cerca és interessant especialment:

- Quan se cerquen temes molt concrets.
- Quan es busca un tema del qual se saben poques dades.
- Quan se cerca una frase o expressió específica.
- Quan es vol obtenir la màxima informació possible sobre un tema.

Exemples de motors de cerca

Exemples de motors de cerca	
Google	http://www.google.com
Alltheweb	http://www.altheweb.com
Altavista	http://www.altavista.com
Wisenut	http://www.wisenut.com
Teoma	http://www.teoma.com
Alexa	http://www.alexa.com
MSN Search	http://search.msn.com

Lectura complementària

En els darrers anys, el mercat dels cercadors d'Internet ha estat objecte d'importants moviments empresarials que han conduït a una progressiva concentració. Un estat de la qüestió ens l'ofereix l'article següent:

J. Grau; J. Guallar (2004). "El negocio de buscar en internet. Análisis del mercado de los buscadores en 2003". *El professional de la información* (vol. 4, núm. 13, pàg. 292-300).

Taula comparativa entre directoris i motors de cerca

Directoris	Motors de cerca
Funcionen per navegació.	Funcionen per interrogació.
La informació s'organitza en categories, sovint arbitràries. A vegades és difícil saber a quina buscar.	És necessari el coneixement del llenguatge d'interrogació i el funcionament de cada motor.
L'actualització no és regular i sol ser lenta.	L'actualització és freqüent.
Tenen criteris d'inclusió poc transparents.	
Apleguen un volum d'informació reduït en comparació dels motors de cerca. No tenen voluntat de ser exhaustius.	Permeten l'accés a una gran quantitat d'informació. Tenen voluntat de ser exhaustius.
Organitzen la informació mitjançant la participació humana seguint procediments intel·lectuals.	Organitzen la informació de manera automatitzada per aplicacions informàtiques.
Els resultats tendeixen a la pertinença.	Els resultats tendeixen a ser exhaustius.

3.3. Metacercadors

Els metacercadors permeten realitzar cerques en diversos cercadors de forma simultània, evitant haver-les de fer per separat a cada cercador i estalviant temps i recursos.

Els metacercadors apleguen la demanda de l'usuari, envien la petició a cada cercador i, posteriorment, reben la informació proporcionada per cadascun dels cercadors i l'ofereixen a l'usuari mitjançant una única interfície.

Malgrat els aparents avantatges, els metacercadors presenten alguns **inconvenients** importants:

- Alguns metacercadors no inclouen els principals cercadors entre la selecció de cercadors on envien la petició de cerca.

- Cada cercador té el seu propi llenguatge d'interrogació i els metacercadors no sempre tradueixen la demanda de l'usuari a la sintaxi específica de cada cercador. Això obliga que les equacions hagin de ser relativament simples.

- No tots els cercadors realitzen control de duplicats.

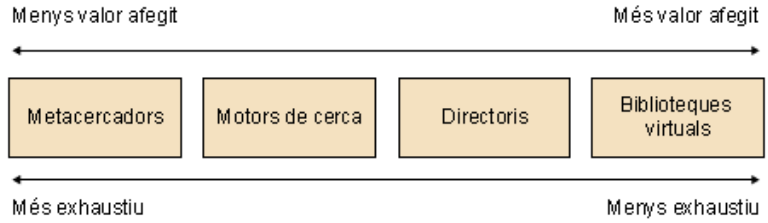
Exemples de metacercadors

Exemples de metacercadors	
Kartoo	http://www.kartoo.com

Exemples de metacercadors	
Metacrawler	http://www.metacrawler.com

Enginys de cerca d'informació

Resum de les principals característiques de les tipologies d'enginys de cerca d'informació analitzats.



3.4. Cercadors de segona generació

La complexitat cada cop més gran que envolta la recuperació de la informació ha conduït a la proliferació de noves eines de cerca que van més enllà de les possibilitats que ofereixen els cercadors tradicionals.

De forma genèrica, aquestes eines, que es coneixen com a **agents de cerca**, són programes informàtics instal·lats als ordinadors dels usuaris capaços de buscar informació al web de forma autònoma d'acord amb uns paràmetres establerts per l'usuari. A més, aporten prestacions complementàries com el filtratge dels resultats, l'eliminació de duplicats...

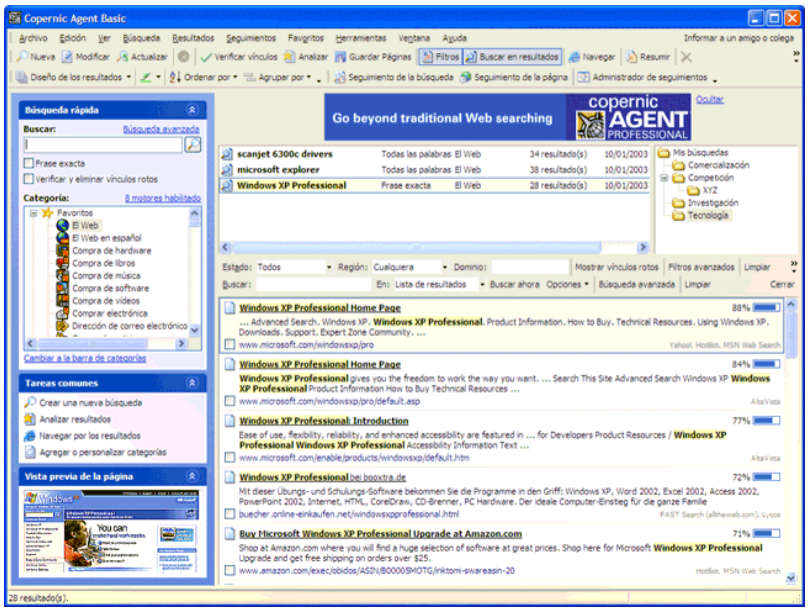
Hi ha una àmplia diversitat d'agents de cerca, però els més coneguts són els anomenats *multicercadors* o *searchbots* (*search robots*). Malgrat que *multicercador* sigui el terme més comú per a referir-se a aquests programes informàtics, no és una terminologia acceptada unànimement. Per alguns autors, *multicercador* s'ha d'usar com a sinònim a *metacercador*, i deixar *searchbots* per a la denominació d'aquests agents de cerca.

Els multicercadors tenen una funció similar als metacercadors, però són programes client instal·lats a l'ordinador de l'usuari que, a més, ofereixen prestacions de valor afegit com les esmentades anteriorment.

Al mercat hi ha un ampli ventall de multicercadors, probablement el més popular és *Copernic*.

Copernic

Es pot obtenir una versió gratuïta de Copernic en l'adreça http://www.copernic.com.



Copernic s'ha convertit en un dels agents de cerca més populars.

Bibliografia

Es pot trobar informació detallada sobre el seu funcionament en l'article: **J. Franganillo; T. M. Figuerola** (2001). "Análisis del buscador múltiple Copernic 2001 Pro". *BiD* (núm. 6). (Data de consulta: 10.09.05) (http://www.ub.es/biblio/dib/06frang.htm).

4. Internet invisible

La cobertura dels cercadors no és exhaustiva. Els motors de cerca que indexen un nombre més gran de pàgines tan sols apleguen una mínima part del total d'informació disponible al web, la gran majoria resta fora de l'abast de les eines de cerca.

Són diversos els motius que expliquen aquest fet, però el principal cal buscar-lo en les mateixes limitacions de la tecnologia que usen els cercadors per a alimentar les seves bases de dades. Els cercadors actuals únicament poden indexar informació provinent de pàgines web estàtiques, i en queden al marge totes les pàgines dinàmiques. És a dir, aquelles pàgines web que integren informació generada després d'una consulta al servidor, primer, per exemple, mitjançant els coneguts CGI-BIN i, més recentment, per mitjà de la tecnologia ASP o PHP.

El conjunt d'informació que resta al marge de l'abast dels cercadors tradicionals rep el nom d'Internet invisible, web profund o infranet. El web profund el constitueixen els continguts de les bases de dades, dels catàlegs, les intranets...

Actualment es considera que la mida del web superficial, el que pot ser gestionat pels cercadors, és d'uns 167 TB. Un terabyte equival a 1.000.000.000.000 de bytes. L'any 2000 s'estimava que el volum d'informació disponible al web era d'entre 20 i 50 TB. El volum calculat per al web invisible és de 91.850 TB.

Bibliografia

P. Lyman; H. R. Varian (2003). "How much information" (document electrònic, data de consulta: 21.3.05).

Per a poder accedir als recursos del web profund, ja existeixen directoris que es dediquen a compilar aquest tipus d'informació, en són exemples:

Recursos per a accedir a Internet invisible	
Internet invisible	http://www.internetinvisible.com
Intelliseek Invisibleweb	http://www.invisibleweb.com
Directe Search	http://gwis2.circ.gwu.edu/-gprice/direct.htm
Infomine	http://infomine.ucr.edu/search.phtml

També alguns agents de cerca, *searchbots*, faciliten la consulta de bases de dades o de catàlegs de forma automàtica i simultània. *BookWhere* n'és un dels exemples més interessants.

Bibliografia

J. A. Ruiz Felipe (2001). "Recuperar información de la internet profunda". *Sociedad de la información* (núm. 3). (Data de consulta: 01.09.05).

L. Vilaragut Llanes; J. R. Carro Suárez. "Para acceder al web profundo: conceptos y herramientas". Comunicació presentada en el Congrés Internacional INFO'2004. (Data de consulta: 01.11.05).

5. Metadades

En els darrers anys, la literatura especialitzada ha dedicat un especial interès a les metadades.

> Estrictament, i de forma literal, les metadades són dades sobre les dades, i el seu interès rau en les possibilitats que aquestes dades ofereixen per a ser usades per a la identificació, descripció i localització de la informació electrònica.

Les metadades troben en la recuperació de la informació la seva principal aplicació. Així, alguns cercadors, com Altavista, van començar a usar les metadades per calcular la rellevància dels documents web. Les metadades, incloses en les etiquetes <META> del llenguatge HTML, aportaven informació diversa sobre el document en qüestió que era usada pel cercador a l'hora d'establir el rànquing de rellevància. Progressivament, el seu ús es va anar abandonant a causa de la generalització del correu brossa amb objectius comercials que alterava els resultats.

Actualment, les metadades, mitjançant l'estàndard RDF, són un dels eixos bàsics del web semàntic, al qual s'ha fet referència breument més amunt.

Lectura recomanada

Una visió crítica del paper de les metadades en la recuperació de la informació a Internet la podem trobar en l'article:

L. Codina (2003). "La web semántica: una visión crítica". *El profesional de la información* (vol. 2, núm. 12, pàg. 149-152).

D'acord amb la definició donada sobre metadades, aquestes no són noves. Els registres d'un catàleg, una descripció bibliogràfica, no deixen de ser metadades, com també ho és el resultat de la catalogació o de la indexació.

Una de les característiques principals de les metadades, i que les diferencia d'altres formes de descripció, és que són els mateixos autors dels documents els responsables de la seva creació. En el cas dels entorns bibliotecaris tradicionals, eren els professionals els que es responsabilitzaven d'aquesta tasca.

Com s'ha comentat més amunt, les metaetiquetes se solen incloure dins de les etiquetes <META> dels documents HTML.

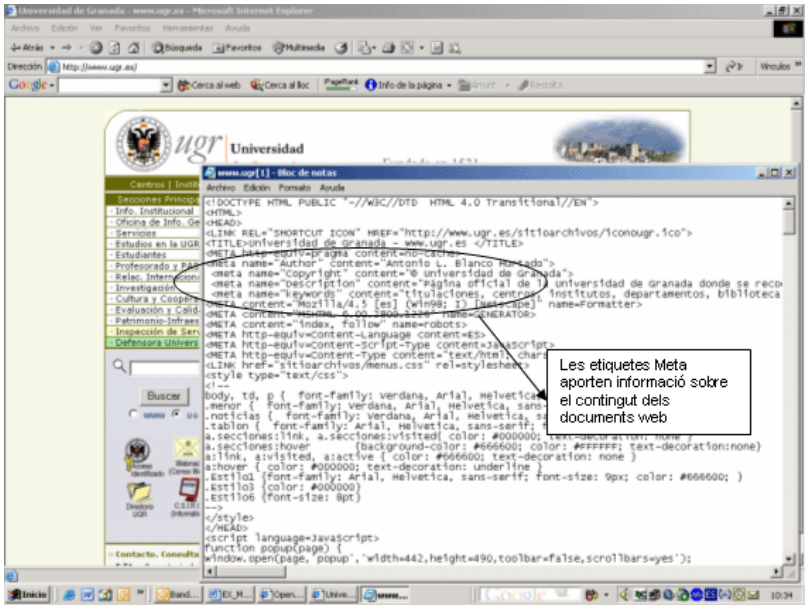
L'etiqueta META

Descrita en la versió 2.0 de l'HTML (http://www.w3.org/MarkUp/html-spec/), l'etiqueta META té com a funcions:

- *to provide a means to discover that the data set exists and how it might be obtained or accessed*;
- *to document the content, quality, and features of a data set, indicating its fitness for use*. (RFC 1866 – www.ietf.org/rfc/rfc1866.txt).

Les metadades poden aportar informació sobre l'estructura i el contingut del document.

Exemple de metadades



No hi ha un únic estàndard per a la descripció de les metadades, tanmateix el ***Dublin Core*** **(DC)** ha adquirit un paper protagonista en l'àmbit de la documentació i, ara com ara, és l'estàndard més estès per a la representació de la informació al web.

El DC neix com un intent d'establir un conjunt bàsic d'elements que permetin la descripció dels documents electrònics de la Xarxa per a la seva cerca i recuperació. El *Dublin Core* està configurat per quinze elements que aporten informació sobre el contingut: títol, paraules clau..., sobre la propietat intel·lectual: autor, editor... i sobre les característiques del document: data, tipus de recurs, format...

Elements del Dublin Core

Els elements de *Dublic Core* permeten aportar informació sobre el contingut, la identitat i els drets d'autor dels documents web. A continuació es mostren els elements que configuren el DC:

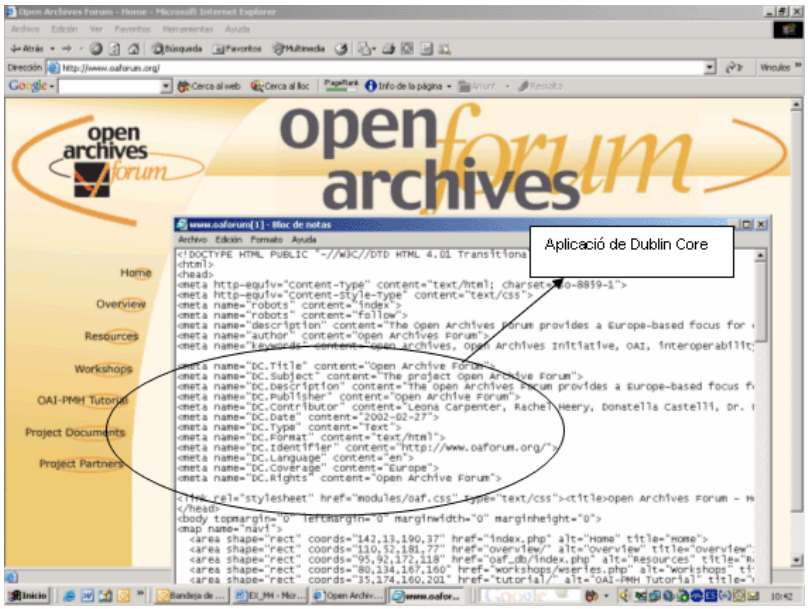
Element Name: Title	
Label:	Title
Definition:	A name given to the resource.

Comment:	Typically, Title will be a name by which the resource is formally known.
Element Name: Creator	
Label:	Creator
Definition:	An entity primarily responsible for making the content of the resource.
Comment:	Examples of Creator include a person, an organization, or a service. Typically, the name of a Creator should be used to indicate the entity.
Element Name: Subject	
Label:	Subject and Keywords
Definition:	A topic of the content of the resource.
Comment:	Typically, Subject will be expressed as keywords, key phrases or classification codes that describe a topic of the resource. Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme.
Element Name: Description	
Label:	Description
Definition:	An account of the content of the resource.
Comment:	Examples of Description include, but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content.
Element Name: Publisher	
Label:	Publisher
Definition:	An entity responsible for making the resource available
Comment:	Examples of Publisher include a person, an organization, or a service. Typically, the name of a Publisher should be used to indicate the entity.
Element Name: Contributor	
Label:	Contributor
Definition:	An entity responsible for making contributions to the content of the resource.
Comment:	Examples of Contributor include a person, an organization, or a service. Typically, the name of a Contributor should be used to indicate the entity.
Element Name: Date	
Label:	Date
Definition:	A date of an event in the lifecycle of the resource.

Comment:	Typically, Date will be associated with the creation or availability of the resource. Recommended best practice for encoding the date value is defined in a profile of ISO 8601 [W3CDTF] and includes (among others) dates of the form YYYY-MM-DD.
Element Name: Type	
Label:	Resource Type
Definition:	The nature or genre of the content of the resource.
Comment:	Type includes terms describing general categories, functions, genres, or aggregation levels for content. Recommended best practice is to select a value from a controlled vocabulary (for example, the DCMI Type Vocabulary [DCT1]). To describe the physical or digital manifestation of the resource, use the FORMAT element.
Element Name: Format	
Label:	Format
Definition:	The physical or digital manifestation of the resource.
Comment:	Typically, Format may include the media-type or dimensions of the resource. Format may be used to identify the software, hardware, or other equipment needed to display or operate the resource. Examples of dimensions include size and duration. Recommended best practice is to select a value from a controlled vocabulary (for example, the list of Internet Media Types [MIME] defining computer media formats).
Element Name: Identifier	
Label:	Resource Identifier
Definition:	An unambiguous reference to the resource within a given context.
Comment:	Recommended best practice is to identify the resource by means of a string or number conforming to a formal identification system. Formal identification systems include but are not limited to the Uniform Resource Identifier (URI) (including the Uniform Resource Locator (URL)), the Digital Object Identifier (DOI) and the International Standard Book Number (ISBN).
Element Name: Source	
Label:	Source
Definition:	A Reference to a resource from which the present resource is derived.
Comment:	The present resource may be derived from the Source resource in whole or in part. Recommended best practice is to identify the referenced resource by means of a string or number conforming to a formal identification system.
Element Name: Language	
Label:	Language
Definition:	A language of the intellectual content of the resource.

Comment:	Recommended best practice is to use RFC 3066 [RFC3066] which, in conjunction with ISO639 [ISO639]), defines two- and three-letter primary language tags with optional subtags. Examples include "en" or "eng" for English, "akk" for Akkadian", and "en-GB" for English used in the United Kingdom.
Element Name: Relation	
Label:	Relation
Definition:	A reference to a related resource.
Comment:	Recommended best practice is to identify the referenced resource by means of a string or number conforming to a formal identification system.
Element Name: Coverage	
Label:	Coverage
Definition:	The extent or scope of the content of the resource.
Comment:	Typically, Coverage will include spatial location (a place name or geographic coordinates), temporal period (a period label, date, or date range) or jurisdiction (such as a named administrative entity). Recommended best practice is to select a value from a controlled vocabulary (for example, the Thesaurus of Geographic Names [TGN]) and to use, where appropriate, named places or time periods in preference to numeric identifiers such as sets of coordinates or date ranges.
Element Name: Rights	
Label:	Rights Management
Definition:	Information about rights held in and over the resource.
Comment:	Typically, Rights will contain a rights management statement for the resource, or reference a service providing such information. Rights information often encompasses Intellectual Property Rights (IPR), Copyright, and various Property Rights. If the Rights element is absent, no assumptions may be made about any rights held in or over the resource.

Tota la informació sobre *Dublin Core* es pot obtenir al web http://dublincore.org. Una versió en castellà de la descripció dels elements està disponible en l'adreça http://www.sedic.es/DCES.pdf.



Exemple d'aplicació de *Dublin Core*

Juntament amb *Dublin Core*, l'altre estàndard en la descripció de documents electrònics és el ***Resource Description Framework* (RDF)**. Estrictament l'RDF no és un model de metadades, com ho és DC, sinó que és una norma per a la creació de models de metadades. RDF es basa en el llenguatge XML (Extensible Markup Language).

Resource Description Framework (RDF)

Les característiques tècniques d'RDF es poden consultar al web del World Wide Web Consortium (W3C): http://www.w3.org/RDF/.

6. Catàlegs

Els catàlegs de biblioteques tenen dues funcions primordials:

- localitzar un document a partir de dades conegudes (p. ex. autor, títol...),
- conèixer quins documents sobre una temàtica determinada formen part de la col·lecció de la biblioteca.

Funcions primordials dels catàlegs

Bàsicament es corresponen amb els dos tipus bàsics de necessitats informatives descrites en el mòdul *Introducció a la cerca i recuperació de la informació*.

La progressiva introducció de la informàtica en la gestió de les biblioteques juntament amb el desenvolupament de les xarxes de telecomunicacions va permetre que les biblioteques posessin a disposició dels usuaris la possibilitat de consultar de forma remota els seus catàlegs. De forma genèrica, els catàlegs automatitzats de les biblioteques es coneixen amb el nom d'**OPAC** (*Online Public Access Catalogue*).

Els primers catàlegs accessibles de forma remota van aparèixer durant la dècada dels vuitanta. Els catàlegs en línia van representar una gran revolució en el món de les biblioteques, ja que van permetre captar usuaris nous i reemplaçar els catàlegs manuals, amb les limitacions que tenien.

En aquell moment l'accés es feia mitjançant una connexió remota Telnet. L'accés via Telnet, tanmateix, es caracteritzava per oferir un entorn molt poc amigable i per obligar l'usuari a conèixer la sintaxi d'interrogació dels llenguatges d'interrogació propis de cada catàleg.

Connexions Telnet

Malgrat que l'accés a catàleg mitjançant connexions Telnet ja pràcticament formi part de la història de la documentació, encara es pot consultar algun catàleg usant aquest protocol.

La irrupció del web a la primeria de la dècada dels noranta va permetre que progressivament els OPAC migressin d'entorns Telnet al nou entorn gràfic. El WWW ofereix interfícies més amigables i intuïtives, que permeten que els usuaris no hagin de conèixer els llenguatges d'interrogació de cada catàleg.

Actualment, els catàlegs de les biblioteques deixen de ser exclusivament bases de dades bibliogràfiques, esdevenint, en alguns casos, veritables bases de dades de text complet. D'aquesta manera, l'usuari no tan sols pot identificar si un document forma part dels fons d'una biblioteca, sinó també accedir-hi i consultar-ne el contingut per mitjà de la pantalla del seu ordinador sense

haver de traslladar-se físicament a la biblioteca. Evidentment això és possible gràcies a la tasca prèvia de digitalització que realitzen biblioteques, organismes públics i empreses.

Bibliografia

Burnett, K.; Bor Ng, K.; Park, S. (1999). "A Comparison of the Two Traditions of Metadata Development". *Journal of American Society for Information Science* (vol. 50, núm. 13. pàg. 1209-1217).

Chowdhury, G. G. (1999). *Introduction to modern information retrieval*. Londres: Library Association Publishing.

Codina, L. (2003). "La web semántica: una visión crítica". *El profesional de la información* (vol. 2, núm. 12, pàg. 149-152).

Koehler, W. (1990). "An Analysis of Web Page and Web Site Consistence and Permanence". *Journal of the American Society for Information Science* (vol. 50, núm. 2, pàg. 162-180).

Méndez, E. (2002). *Metadatos y recuperación de la información*. Gijón: Trea.

Ruiz Felipe, J. A. (2001). "Recuperar información en la internet profunda". *Sociedad de la información* (núm. 3). (Data de consulta: 01.09.05).

Vilaragut Llanes, L.; Carro Suárez, J. R. (2004). "Para acceder al web profundo: conceptos y herramientas". Comunicació presentada en el Congrés Internacional INFO'2004. (Data de consulta: 01.11.05).