Sistemes de gestió documental i bases de dades documentals

Ernest Abadal Falgueras
Lluís Codina Bonilla

PID_00168110

Universitat Oberta de Catalunya
www.uoc.edu

Índex

Introducció

Les bases de dades són la millor tecnologia de què disposem en l'actualitat per a gestionar informació, ja que és l'únic sistema que permet processar la informació d'una manera que és, alhora, segura, ràpida i eficaç. De fet, hi ha altres tecnologies basades en ordinadors per a gestionar informació. Per esmentar-ne algunes: editors de text, programes de fulls de càlcul, gestors de fitxers, navegadors d'Internet, etc. Però només les bases de dades permeten accedir a la informació selectivament, mostrar-la d'una manera diferent a diferents grups d'usuaris, explotar-la d'una manera diferent si canvien els objectius, etc., i tot això en el marc d'una seguretat i una confidencialitat relatives tant enfront d'accessos maliciosos com enfront d'errors involuntaris.

En aquest context, les bases de dades documentals compleixen una funció d'una importància particular. Es tracta d'un tipus de producte pensat per a tractar no tant amb dades sinó amb informació cognitiva o coneixement. Per explicar aquesta funció cal tenir en compte que, quan parlem d'informació, podem estar pensant en les dades d'una factura (qui l'emet, per quin import, qui l'ha d'abonar, etc.) o d'una tesi doctoral, per esmentar dos extrems.

En l'apartat "Alguns conceptes bàsics" del mòdul "Sistemes de bases de dades" hi ha un debat sobre els conceptes dades, *informació* i *coneixement*.

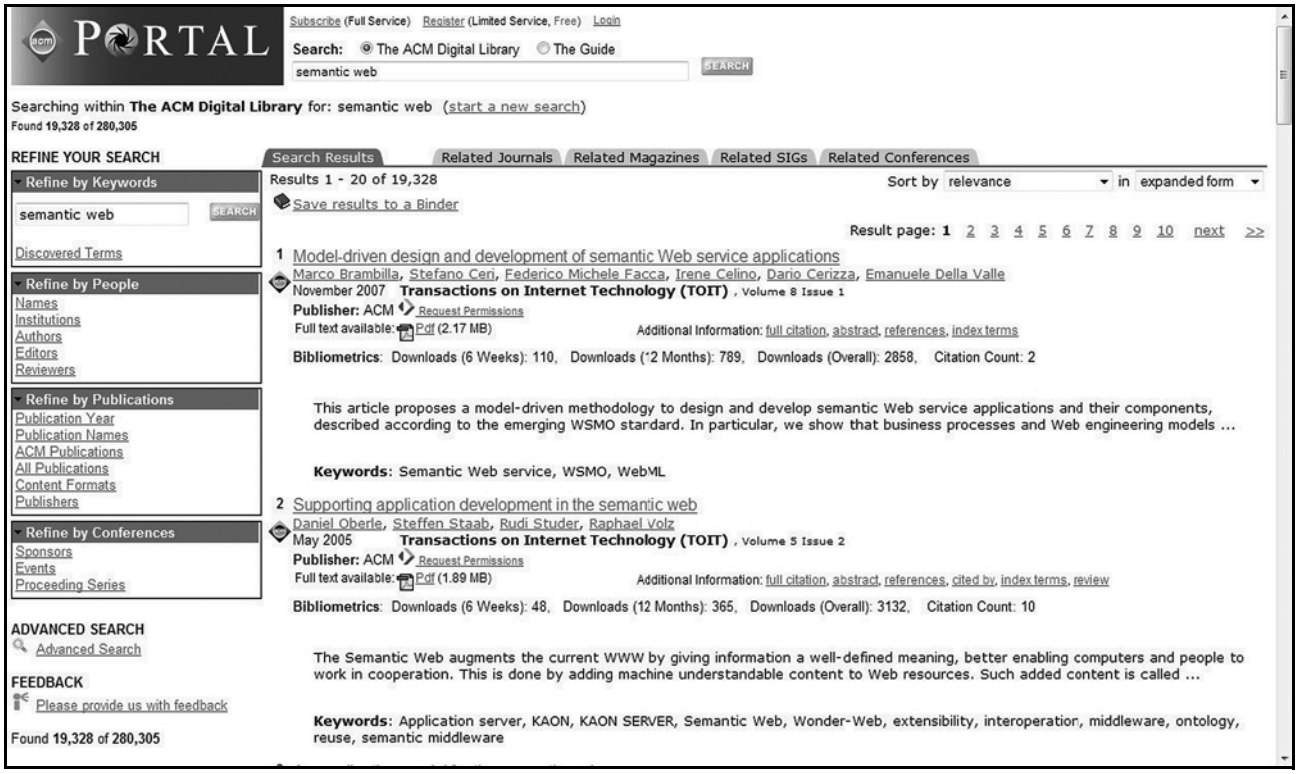
En el primer exemple estem parlant d'informació administrativa, mentre que en el segon estem parlant d'informació cognitiva, de coneixement expressat i registrat en un document (en aquest cas una tesi). En el primer document (la factura) hi ha dades numèriques que són fàcils de representar, per exemple, en forma de taula amb valors atòmics (cada cel·la un valor únic). També hi ha text, però en forma de dades factuals breus i compactes (un nom propi, una adreça, un nom de producte, etc). En el segon document, es poden trobar dades factuals però sobretot hi ha text en forma de discurs raonat, exposició de teories, raonaments inductius o deductius, etc. El contingut d'aquest segon tipus de documents no pot ser reduït a una taula amb valors atòmics.

Aquest és un dels motius pels quals els sistemes de gestió de bases de dades relacionals, basats en taules amb valors atòmics, no puguin gestionar bé documents cognitius com els esmentats. Altres exemples d'aquests documents són els articles de revista, els informes tècnics o científics de qualsevol tipus, les informacions periodístiques, la documentació de manteniment d'equips, les patents, etc.

De fet, la tecnologia que fonamenta les bases de dades documentals és l'única que pot donar suport a aplicacions tan importants com les bases de dades científiques o acadèmiques (Web of Knowledge, Scopus, etc.), els cercadors d'Internet, les hemeroteques i els repositoris digitals de la web, els cercadors

interns de llocs i Intranets, els catàlegs de biblioteques, els portals de revistes, les bases de dades de patents, de tesis doctorals, etc.

Figura 1. Pàgina de resultats d'una típica base de dades documental de tipus acadèmic (ACM, en aquest cas)



Si mirem enrere, podem veure que, històricament, les primeres bases de dades documentals sorgeixen a finals dels anys seixanta als EUA i apareixen vinculades tant al món de la informació periodística com al de la informació científicotècnica. Des de llavors no han deixat d'estendre's a altres terrenys i activitats socials, com hem intentat explicar en el paràgraf precedent.

Amb l'objectiu de contribuir a fixar els conceptes bàsics que manejarem en els apartats següents, introduirem la dicotomia *base de dades* contra *sistema de gestió de bases de dades*, que no sempre es diferencia amb claredat encara que es tracta d'una distinció de gran importància. En primer lloc, cal recordar que una *base de dades* és un conjunt o una col·lecció de dades estructurades emmagatzemades digitalment, mentre que un *sistema de gestió de bases de dades* (SGBD) és el programa que permet la creació, el manteniment i l'explotació de la base de dades.

Els conceptes *base de dades* i *SGBD* es defineixen en el mòdul "Sistemes de bases de dades".

A partir d'aquí podem comentar les característiques principals de les bases de dades.

1) Les dades estan interrelacionades i estructurades seguint un model

Les dades han de posseir alguna estructuració interna, és a dir, no es pot tractar d'un mer dipòsit o magatzem d'informació. Per a això cal recórrer a diversos models que ajuden a estructurar i interrelacionar les dades per a facilitar la recuperació de la informació.

Els models en la fase de disseny es tracten en el subapartat "La fase de disseny" de l'apartat 3.

2) Les dades estan emmagatzemades en un suport informàtic

Aquest és un altre aspecte fonamental: el contingut d'una base de dades ha d'estar gravat en un suport digital. D'una altra manera ens trobem, per exemple, davant d'un llistat imprès.

3) Existeix un programa que s'ocupa de la gestió i manipulació de les dades

Els sistemes de gestió de bases de dades (SGBD) són els programes que permeten crear les bases de dades, accedir-hi i manipular-les. Sense el seu concurs no es podria donar sortida al que constitueix l'objectiu principal d'una base de dades: la selecció, recuperació i explotació de la informació que conté.

4) Les dades seran usades o bé per altres programes informàtics o bé per persones

En la concepció característica de la informàtica de gestió, les bases de dades amb freqüència no són per a usuaris finals (persones), sinó per a donar suport a processos informàtics que duen a terme programes d'ordinador. Per exemple, els continguts d'una base de dades de recursos humans serviran, principalment, per a confeccionar de manera automàtica la nòmina de cada mes. En canvi, les bases de dades de tipus documental gairebé sempre estan orientades a donar servei a usuaris finals: per exemple, als usuaris d'un centre de documentació o d'una biblioteca.

Després d'aquesta presentació bàsica, aprofundirem en les bases de dades documentals a partir de les diferents operacions que s'hi poden fer:

- **Producció i administració** (apartat 1), en què ens ocuparem fonamentalment de l'estructura i les característiques dels programes informàtics per a crear bases de dades documentals.

- **Distribució** (apartat 2), en què introduirem el concepte *interfície de consulta d'una base de dades documental* i es donaran indicacions per a avaluar-la i dissenyar-la.

- **Disseny** (apartat 3), en què ens centrarem en la metodologia per a la creació de bases de dades documentals.

- **Avaluació** (apartat 4), en què s'oferiran els indicadors fonamentals per a l'avaluació d'una base de dades documental.

Objectius

L'estudi d'aquest mòdul us permetrà conèixer a fons els continguts següents:

1. Saber quines són les característiques dels sistemes de gestió de bases de dades documentals.

2. Saber quins tipus de bases documentals hi ha.

3. Conèixer com es pot crear una base de dades documental.

4. Saber quin és el disseny més adequat d'una base de dades documental.

1. Producció i administració de bases de dades

L'element fonamental per a la producció i l'administració de bases de dades són els sistemes de gestió de bases de dades (SGBD), els programes informàtics que permeten la creació i l'explotació de bases de dades.

En aquest apartat analitzarem les característiques generals dels SGBD, la seva tipologia (SGBDR i SGD) i ens centrarem fonamentalment a descriure l'estructura, el funcionament i el mercat dels diferents tipus d'SGD (SGBDD i sistemes d'indexació).

1.1. Els sistemes de gestió de bases de dades (SGBD)

Per a aprofundir en la comprensió dels SGBD podem prendre com a referència la definició següent:

> "[Un SGBD es un] conjunto coordinado de programas, procedimientos, lenguajes, etc. que suministra a los diferentes tipos de usuario los medios necesarios para describir y manipular los datos almacenados en la base de datos, garantizando su seguridad." (Miguel, 1997, pàg. 38)

Com ja hem apuntat en la introducció, no tots els SGBD són iguals en funcions i objectius i podem distingir entre els **SGBD relacionals** i els **SGBD documentals** (o textuals). Vegem-ne les diferències:

1) Sistemes de gestió de bases de dades relacionals (SGBDR)

Solen utilitzar un model lògic de dades denominat *relacional*. Són programes especialment adequats per a la gestió d'informació molt estructurada (dades pròpiament dites). En la concepció informàtica clàssica, de fet, és l'únic tipus de sistema de gestió de bases de dades que es considera. Estan molt implantats en l'àmbit de l'empresa per a gestionar i automatitzar processos, de manera que moltes bases de dades gestionades amb SGBDR no estan pensades per a ser consultades per persones (usuaris), sinó per a ser usades com a part de processos informàtics (generar la facturació mensual, per exemple).

Exemple

S'utilitzen per a la gestió de volum de vendes, sous o existències de magatzem, etc.

2) Sistemes de gestió documentals (SGD)

Solen utilitzar un model lògic denominat *textual*. La seva característica comuna és que estan concebuts per a gestionar la classe d'informació amb gran quantitat de text de tipus discursiu i poc estructurat (des del punt de vista informàtic) que és típica dels documents cognitius. Mentre que un dels ele-

Exemple

S'utilitzen per a la gestió d'articles de revistes, pàgines web o reportatges fotogràfics, per esmentar-ne tres exemples molt diferents.

ments fonamentals del model relacional són les taules homogènies (files i columnes iguals), en el cas del model textual ho són el registre irrestricte (sense limitacions) i els índexs analítics.

Sobre el registre irrestricte i els índexs analítics, vegeu el subapartat 1.2.1.

La taula següent ofereix un resum dels trets diferencials fonamentals dels dos grans tipus d'SGBD considerats. En els apartats següents ens centrarem en els SGBD documentals, que són els utilitzats per a la creació i l'explotació de les bases de dades documentals.

Taula 1. Sistema relacional contra sistema documental

Tipus de sistema	Context	Tipus de dades	Finalitat
Relacional	Gestió administrativa, comptable, etc., típica de qualsevol organització pública o privada.	Estructurat i molt regular (p. ex. xifres de vendes, o adreces postals).	Gestió, administració, supervisió, planificació, etc., d'empreses i tot tipus d'organitzacions.
Documental	Adquisició de coneixement i satisfacció de necessitats d'informació més o menys complexes.	Text de tipus discursiu, propi d'articles de revistes, notícies de premsa, etc., o text descriptiu per a *descriure* objectes multimèdia: imatges, vídeo, so, etc.	Estudi, investigació i adquisició de coneixements al servei de projectes, processos d'ensenyament-aprenentatge, investigació, suport a la R+D, etc.

Què impedeix usar un sistema relacional (SGBDR) per a la gestió documental? En principi, res. És a dir, res no ho impedeix si el volum d'informació a tractar és petit, si no necessitem prestacions de control terminològic i si no necessitem sortides en formats bibliogràfics específics, per esmentar només tres grups de prestacions funcionals.

Vegem per separat els dos primers aspectes que, probablement, són els fonamentals. Els SGBDR no indexen tot el contingut dels camps de text. Per defecte, els camps amb molta informació textual o bé no s'indexen o bé indexen únicament la primera paraula de cada camp. En aquest context, si la base de dades conté poca informació i s'utilitza un ordinador suficientment ràpid, una cerca seqüencial pot imitar l'ús d'un índex de tipus documental. No obstant això, quan creixi la base de dades, les prestacions del sistema es degradaran. L'experiència indica que, a partir d'un determinat nombre de documents, un sistema relacional difícilment podrà gestionar amb eficàcia un contingut de tipus cognitiu. En canvi, el mateix sistema relacional podrà gestionar amb gran eficàcia milions de registres de tipus tabular (és a dir, dades de l'estil d'adreces, comptabilitat, dades de vendes, etc.).

Indexar (elaborar un índex)

És l'acció de registrar dades ordenadament per a obtenir resultats rellevants de manera més ràpida en una cerca d'informació.

En segon lloc, amb **prestacions de control terminològic** ens referim a la possibilitat de definir i utilitzar diccionaris de paraules buides, diccionaris de sinònims, tesaurus, etc., que controlen el resultat de la indexació i faciliten la realització de cerques. Altres deficiències dels sistemes relacionals pel que fa a la gestió documental es refereixen a dificultats tècniques per definir el nombre òptim de caràcters de cada camp (que cal prefixar per endavant), el nombre

òptim de camps que s'utilitzaran per a contenir descriptors, l'absència d'eines per a gestionar i produir bibliografies, etc.

Així, podem reprendre la pregunta anterior "què impedeix utilitzar un sistema relacional per a la gestió documental?" i respondre-la ara: tot. Tot ho impedeix si el que necessitem és gestionar el contingut de grans volums de documents de tipus cognitiu o si necessitem utilitzar algun tipus de control terminològic per a optimitzar els resultats.

1.2. Sistemes de gestió documental (SGD)

Els sistemes de gestió documental, que en anglès reben denominacions com *information retrieval systems, text retrieval systems*, *document retrieval systems* (o *digital asset management* quan s'utilitzen per a documents icònics), són el tipus de programa especialment adequat per a la gestió d'informació textual i de documents cognitius.

Com ja hem indicat, en general, els SGD estan concebuts per a gestionar documents de tipus científic (articles, ponències, tesis, etc.), tècnic (informes, patents, etc.) o cultural (articles de premsa, fotografies, etc.). Permeten, per tant, la gestió de fons documentals de qualsevol naturalesa i això inclou la gestió de qualsevol col·lecció de textos, imatges i objectes multimèdia (so, música, vídeo, etc). El model aproximadament similar (però no idèntic) que segueix la majoria dels SGD se sol denominar, a falta d'un nom millor, *model textual*.

1.2.1. Característiques del model textual

El model textual o documental presenta, almenys, cinc característiques importants:

1) Un model de registre irrestricte

En els SGD no hi ha restriccions prèvies al tipus de registre que poden manejar. En aquest sentit, els models de registre poden anar des d'esquemes totalment oberts, com si es tractés de documents d'un editor de text (per exemple, askSam), fins a models perfectament articulats en camps i tipus de dades (per exemple, Inmagic, CDS/ISIS), passant per tipus intermedis que aporten una bona flexibilitat per a treballar amb camps articulats, però sense complicacions excessives (per exemple, File Maker o Inmagic). El model irrestricte es refereix també a la possibilitat que en una mateixa base de dades puguin conviure models de registres diferents (CDS/ISIS, Inmagic o RefWorks).

Figura 2. Exemple d'un registre documental típic (en aquest cas, del programa RefWorks)

No. de Identificación:	560
Tipo de Referencia:	Artículo de Revista Acadêmica (Journal)
Tipo de fuente:	Impreso
Idioma de salida:	Inglés
Autores:	Diaz Noci,Javier
Título:	Multimedia and Reading Ways: a State of the Art
Publicación Completa:	COMUNICAR
Año de Publicación:	2009
Fecha de Publicación - Formato Libre:	OCT
Ejemplar:	33, Sp. Iss. SI
Página Inicial:	213
Otras Páginas:	219
Descriptores:	Multimedia; reading; online journalism; communication; hypertext; interactivity}
Resumen:	Multimedia is one of the less studied characteristics, probably because of the less-developed level of the digital language. Along with hypertext and interactivity, it is one of the characteristics that defines the digital edition. Those characteristics have been always studied from the point of view of production, although not so much from the point of view of reception. How do users read a digital text? The reader's participation, reading depth, different trailblazing, the relation user-interface and the conception of multimedia text as a module of a database introduce major changes in the reception of the text, which can and must be studied.
Notas:	Article}
Editorial:	GRUPO COMUNICAR
Lugar de Publicacion:	APDO CORREOS 527, HUELVA, 21080, SPAIN}
ISSN/ISBN:	1134-3478
Dirección/Afiliación	Noci, JD (Reprint Author), Univ Pompeu Fabra Barcelona, Fac Comunicac, Dept Comunicac, Barcelona, Spain. Univ Pompeu Fabra Barcelona, Fac Comunicac, Dept

2) Capacitat monobase o multibase indistintament

És característic d'alguns SGD que únicament puguin obrir i fer servir una sola base de dades cada vegada (askSam). No obstant això, cada vegada són més els SGD amb capacitat *multibase*, és a dir, que poden obrir i fer servir més d'una base de dades alhora (CDS/ISIS, Inmagic, o File Maker).

3) Índex analític (fitxer invers)

El fitxer invers és un índex (o un conjunt d'índexs) compost per totes i cadascuna de les paraules que apareixen en tots i cadascun dels registres de la base de dades. Des del moment en què aquestes paraules representen temes, idees i conceptes, l'índex d'una base de dades documental és una representació de tots els assumptes presents en tots els documents que formen part de la base de dades. Els índexs analítics se solen basar en una estructura denominada *fitxer invers* o *fitxer invertit*.

L'estructura dels índexs analítics està optimitzada per tal de permetre l'existència de valors repetits (documents indexats amb el mateix descriptor), per a fer cerques en documents de text complet amb gran rapidesa i per a dur a terme tasques de control terminològic.

Exemple

Si en una base de dades documental apareix cent vegades el terme "economia", per contra, hi ha una sola entrada en el fitxer invertit (en l'índex d'un sistema relacional hi hauria d'haver cent entrades). Els fitxers invertits relacionen a més dades de context amb cada terme de l'entrada, per exemple, la seva freqüència, la seva posició exacta en cada registre (nombre d'ordre), els possibles sinònims, etc.

En la classe d'índexs analítics que permeten els fitxers invertits, cada terme o entrada de l'índex és únic, la qual cosa facilita temps de resposta molt baixos.

En les dues taules següents s'il·lustra el concepte d'un índex analític mitjançant el sistema del fitxer invertit.

Taula 2. Composició típica d'un índex invertit

Element	Explicació
Terme	Totes i cadascuna de les paraules que formen part dels registres o dels documents de la base de dades (i que no consten en el fitxer de paraules buides). Sempre són termes únics, és a dir, hi ha una sola entrada per a cada terme encara que aparegui moltes vegades en un o en molts registres de la base de dades.
Freqüència	Nombre de registres (per tant, nombre de documents) en els quals apareix el terme. En alguns fitxers invertits també es consigna el nombre total de vegades (freqüència) en què el terme apareix.
Localització	Indicació dels paràmetres de localització, imprescindible per a la recuperació. La informació necessària consta, almenys, dels elements següents: número document – número de camp (si és que hi ha camps) – número de paraula. El motiu és que cal conèixer la posició absoluta de la paraula en el document per a poder aplicar correctament alguns operadors com el de proximitat.

Fitxer de paraules buides

Conté paraules sense significat semàntic (com articles, pronoms, preposicions…) que són transparents als motors de cerca (no les indexen) i actua de filtre per impedir que formin part del fitxer invers.

Taula 3. Exemple d'índex invertit

Terme	Freqüència	Localització
…	…	…
Barcelona	2	(00017, 03, 01) (03401, 01, 04)
…	…	…
Madrid	2	(00017, 03, 03) (17200, 02, 01)
…	….	…
Saragossa	3	(00017, 03, 04) (03401, 01, 02) (17001, 04, 01)
…	…	…

En l'exemple d'índex de la taula 3 s'inclouen, per simplificar, tan sols tres entrades de l'índex: les corresponents a les paraules *Barcelona*, *Madrid*, *Saragossa*. Si mirem l'entrada *Barcelona*, per exemple, veiem que en total hi ha dos registres en la base de dades que contenen la paraula *Barcelona* (vegeu la columna Freqüència). Com que hi ha dos registres amb la paraula *Barcelona*, en la columna Localització veiem dos vectors, és a dir, dos conjunts de dades: (00017, 03, 01) i (03401, 01, 04). Segons aquests vectors, els dos registres que contenen la paraula *Barcelona* són el 00017 i el 03401, és a dir, el primer de cadascun dels tres nombres que formen cada vector (**00017**, 03, 01) i (**03401**, 01, 04).

Diem que els conjunts (00017, 03, 01) i (03401, 01, 04) són vectors perquè en aquests conjunts la posició de cada element és significativa. D'aquesta manera, el primer nombre sempre és l'identificador del registre, el segon nombre és l'identificador del camp i el tercer nombre identifica el número d'ordre de la paraula en qüestió dins del camp considerat. Això significa que l'índex invertit del nostre exemple correspon a una base de dades amb un model de registre com aquest:

01	Títol
02	Autor
03	Font
04	Descriptors
…	…

En concret, veiem que *Barcelona* apareix en el camp número 3 del primer registre (00017, **03**, 01) però, en canvi, apareix en el camp número 1 del segon registre (03401, **01**, 04). Per tant, això significa que en el registre número 0017 la paraula *Barcelona* apareix en la **Font** (camp 03) i, en canvi, en el registre número 03401 apareix en el camp **Títol** (camp 01).

Així mateix, veiem que la paraula *Barcelona* apareix en primera posició en el primer dels dos registres (00017, 03, **01**), però apareix en quarta posició en el segon registre (03401, 01, **04**), etc.

Per acabar d'entendre com un SGD genera (i interpreta en el moment de la cerca) l'índex anterior, representarem com podria ser el registre que correspondria al segon vector [(03401, 01, 04)] de la paraula *Barcelona*:

ID Camp		**03401**
01	Títol	Història il·lustrada de **Barcelona**
02	Autor	F. Pujol
03	Font	Vic: Editorial ZYX, 2010
04	Descriptors	Barcelona, Història

Si comparem el registre anterior amb el vector corresponent [(03401, 01, 04)] podem veure la correspondència d'una manera clara: el primer nombre del vector és el número del registre (**03401**), el segon nombre és l'identificador del camp (**01**, per tant, el **Títol**) i el tercer és l'ordre de la paraula en qüestió en la frase (la quarta paraula en el títol del document).

4) Eines de control terminològic o lingüístic

Encara que hi ha grans diferències entre ells, gairebé tots els SGD solen disposar de diverses eines de control terminològic. La més simple és la possibilitat d'utilitzar diccionaris de paraules buides, és a dir, de termes que no s'usaran per a indexar els documents. La més sofisticada és la possibilitat d'usar un o més tesaurus, és a dir, un llenguatge documental que permet establir relacions lògiques entre els termes i els descriptors d'una base de dades. Entremig, hi ha diverses possibilitats: ús de sinònims, llistes de descriptors, etc.

Diccionari de paraules buides

Stoplist és un fitxer de paraules buides.

Tesaurus

És un diccionari estructurat de conceptes (amb jerarquies i relacions).

5) Llenguatge i interfícies de consultes orientats a l'usuari

Els SGD estan orientats a l'usuari i no tant a altres programes informàtics. Per això el seu llenguatge d'interrogació disposa d'eines que faciliten la conversió d'una necessitat d'informació de l'usuari en una estratègia de consulta, així com facilitats per al manteniment i la gestió d'operacions de cerca complexes, que poden requerir consultes reiterades. Les possibilitats i prestacions en

aquest sentit són molt més grans i més versàtils que les que ens ofereixen els SGBDR i això s'explica, fonamentalment, per dues raons: en primer lloc, perquè el tipus d'informació que contenen és diferent i, en segon lloc, perquè, com ja hem vist anteriorment, les necessitats dels usuaris d'aquest tipus de sistemes són molt diferents dels usuaris de sistemes administratius.

1.2.2. Síntesi: Sistema relacional contra sistema documental

Les diferències comentades entre sistemes relacionals i documentals les sintetitzem i sistematitzem en la taula 4. La comparació s'ha dut a terme partint de dos models purs als quals segurament no tots els programes tenen per què ajustar-se. Per exemple, alguns SGD estan incorporant eines que permeten relacionar bases de dades com si fossin relacionals (Inmagic, FileMaker). A més, alguns programes relacionals integren sota una mateixa interfície o capa de programació un sistema relacional i un sistema documental (Oracle o BRS).

D'aquesta manera, la tendència que se segueix va cap a la integració gradual de les prestacions d'un model i de l'altre en un sol programa. Així doncs, en el mercat podem trobar programes que, malgrat que pertanyen a un dels tipus, tenen algunes característiques de l'altre. No obstant això, és útil comparar els models "purs" de cada categoria.

Taula 4. Diferències principals entre SGBDR i SGD

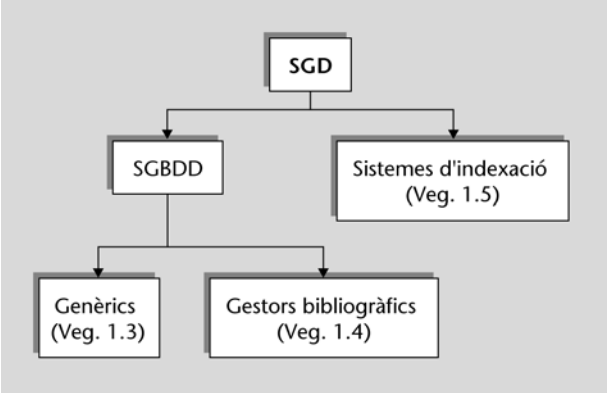
SGBDR	SGD
Estructura	
• Tabular (taules per a representar dades). • Camps amb longitud fixa. • No hi pot haver grups de repetició. • Taules homogènies.	• Textual (model irrestricte). • Camps i registre de longitud variable. • Camps repetibles. • Es poden combinar estructures diferents dins de la mateixa base de dades (per a representar diversos tipus de document, per exemple).
• Conjunt de diverses taules, amb la possibilitat de crear taules noves mitjançant operacions d'àlgebra relacional integrades en el llenguatge de consulta del sistema de gestió de la base de dades.	• Monobase (fitxer pla) o bé amb un sol tipus de registres per cada base de dades (Knosys) o bé amb diversos tipus de registres en la mateixa base de dades (askSam), però podent obrir i fer servir una sola base de dades cada vegada. • Multibase, o bé amb la possibilitat d'obrir i consultar dades de diverses bases de dades alhora (Inmagic), o bé amb la possibilitat de relacionar i operar amb diverses bases de dades alhora en un estil similar al relacional (CDS/ISIS o Inmagic).
• No utilitzen índexs analítics (fitxer invers).	• Usen índexs analítics (fitxer invers).
• Instruments de recuperació (recuperació *determinista*) limitats.	• Instruments de consulta i recuperació amplis, amb moltes ajudes per a les cerques: cerca global en qualsevol camp, operadors booleans, de proximitat, combinació de conjunts de cerca, consulta d'índexs, etc. (recuperació *probabilista*).

SGBDR	SGD
Estructura	
• No disposen de controls terminològics.	• Disposen d'instruments de control terminològic per a la indexació, per a l'entrada de dades i per a la consulta (paraules buides, llistat d'autoritats, sinònims, etc.).
Objecte	
• Informació molt estructurada (informació de gestió, administrativa, etc.).	• Informació poc o gens estructurada (documents científics, tècnics o culturals, o bé documents icònics).
• Informació molt volàtil: les dades acostumen a canviar amb freqüència.	• Informació acumulada: les dades solen ser permanents i acumulatives.
Àmbit	
• Gestió administrativa (ofimàtica). Per exemple, matriculacions, nòmines, etc.	• Serveis d'informació i documentació (centres de documentació, biblioteques, museus, editorials, bancs d'imatges, etc.).

1.2.3. Tipologia de SGD

Pel que fa a la tipologia de *programes de gestió documental*, podríem resumir en dos els models principals presents actualment en el mercat:

Figura 3. Tipologia d'SGD



1) Sistemes de gestió de bases de dades documentals (SGBDD)

Els primers programes de gestió documental estaven pensats per gestionar solament referències de documents, i no el text del document complet. Alguns exemples són CDS/ISIS o Inmagic.

SGBDD genèrics

En l'apartat 1.3. es tracten amb detall.

També s'han de considerar els *sistemes de gestió bibliogràfica*, com una variant dirigida a usuaris personals. Els exemples més coneguts són RefWorks, EndNote, Zotero o ProCite.

Els sistemes de gestió bibliogràfica

Es tractaran amb més detall en l'apartat 1.4.

2) Sistemes d'indexació

Aquests sistemes estan especialment orientats al tractament del text complet dels documents. Són programes que no necessiten definir models de registre, encara que alguns d'ells ho poden fer de manera opcional. La seva especificitat rau en la capacitat de generar índexs analítics (fitxers invertits) del contingut dels documents i desar-los en un disc dur o en una xarxa de discos durs. Els documents continuen estant en el seu format original i l'índex és únicament un punter a cada document concret. Per visualitzar el document, el sistema activa al programa amb el qual va ser creat el document en cada cas. Es denominen **sistemes d'indexació o motors de cerca**. Alguns exemples de sistemes d'indexació o motors de cerca són Autonomy, o Google Search.

Sistemes d'indexació o motors de cerca

Es tracten amb més profunditat en l'apartat 1.5.

Per tal de diferenciar entre tots dos models d'SGD cal observar quines són les funcions prioritzades. En el cas dels SGBDD destaca especialment l'apartat de definició de la base de dades, que facilita l'aplicació d'un diccionari de dades complex, mentre que els sistemes d'indexació tenen molt desenvolupat el mòdul d'indexació, que permet generar els índexs invertits de documents extensos.

1.3. Sistemes de gestió de bases de dades documentals (SGBDD)

Sota aquesta denominació s'inclourien els primers SGD, els més tradicionals, aquells que faciliten bàsicament la gestió de referències de documents de tota mena. Aquests programes comparteixen una sèrie d'elements estructurals que permeten la creació i explotació de bases de dades.

1.3.1. Estructura

Tenen cinc mòduls bàsics:

1) Definició dels registres de la base de dades

Aquest grup funcional està relacionat amb el disseny i la creació de les bases de dades. En particular, permet definir camps, especificar el seu comportament i definir models de registres mitjançant agrupacions de camps. En el procés de creació de bases de dades, les especificacions detallades de cada model de registre i del comportament de cada camp se solen detallar prèviament en un document escrit que rep el nom de *diccionari de dades* (vegeu 3.3.1.). En el diccionari de dades i en el mòdul funcional corresponent del SGBDD també es detallen aspectes relacionats amb el tipus de dada assignada a cada camp (textual, numèrica, lògica, etc.) i amb el control terminològic (camp amb descriptors, camp indexat, etc.).

2) Manteniment

El mòdul de manteniment controla les operacions d'altes, baixes i modificacions de registres.

3) Indexació i recuperació

Es tracta d'un dels mòduls fonamentals i, sens dubte, el més específic d'aquest tipus de programes. Aquí s'inclouen les funcions relacionades amb el procés de generació dels índexs, les prestacions de recuperació (operadors booleans, de proximitat, etc.) les maneres de mostrar i oferir els resultats als usuaris.

4) Sortida i intercanvi

Aquestes opcions comprenen tant els aspectes relacionats amb la sortida dels registres (exportacions) com les operacions que s'ocupen de la incorporació i adaptació de fitxers externs de registres (importacions). Tots dos processos són fonamentals en un SGBDD, ja que asseguren la difusió i la possibilitat d'intercanvi de les seves dades amb l'exterior del sistema.

5) Administració de la base de dades

Aquest mòdul agrupa totes les funcions i tots els processos relacionats amb el control i la gestió de la base de dades, és a dir, el sistema de seguretat (poder crear grups d'usuaris i l'adscripció de diferents privilegis a cada grup d'usuaris, així com l'administració de noms d'usuari i de contrasenyes), i la programació i les modificacions en la interfície.

1.3.2. Mercat

A continuació es presenta una breu fitxa descriptiva individualitzada dels principals SGBDD que es poden trobar actualment en el mercat espanyol. Hem reduït la llista a quatre programes que considerem que són els que estan més implantats i que poden respondre a un tipus de necessitat de més alt nivell (seria el cas d'Inmagic DB/Text i de CDS/ISIS) o d'un nivell mitjà (en aquesta situació es troben FileMaker i Knosys).

Nom	CDS/ISIS
Productor	Unesco <http://www.unesco.org/webworld/isis/isis.htm>
Distribuïdor	Cindoc <http://www.cindoc.csic.es>
Comentaris	• Especialment adequat per al tractament de la informació bibliogràfica. • Utilització de subcamps. • Indexació amb diverses tècniques (paraula a paraula, grups de paraules, camp sencer, subcamps).
Exemples	• Icomos: http://databases.unesco.org/icomos/ • Biblioteca de l'IDAIC: http://www.dba.it/idaic/ricerca.html • Bases de dades de Bireme: http://bases.bvs.br

Nom	FileMaker
Productor	Claris <http://www.filemaker.fr/spain>
Distribuïdor	Claris <http://www.filemaker.es>
Característiques	• D'ús molt fàcil. • Molt versàtil, de fet és el programa amb capacitat documental més integrat alhora en el món ofimàtic. • Capacitat relacional. • Eines de control terminològic escasses.
Exemple	Bibliografia sobre biblioteques públiques: http://www.bibliotecaspublicas.info/

Nom	Inmagic
Productor	Inmagic <http://www.inmagic.com>
Distribuïdor	Doc 6 <http://www.doc6.es>
Característiques	• Molt versàtil: adequat per al tractament de referències bibliogràfiques i per a gestionar qualsevol tipus d'objecte o entitat. • Capacitat relacional. • Àmplies possibilitats d'adaptació i personalització de les interfícies d'usuari. • Àmplies possibilitats de control terminològic. • Gestió integrada de tesaurus. • Dues formes diferents d'indexació (paraula a paraula, per frases).
Exemples	• Foundation center: http://lnps.fdncenter.org/search.html? • Directori de bases de dades: http://www.andornot.com/webpublinks

Nom	Knosys
Productor	Micronet <http://www.micronet.es>
Distribuïdor	Micronet
Comentaris	• Utilització fàcil. • Possibilitats limitades de control terminològic. • No diferencia entre el fitxer de definició de camps, l'entrada de dades i la visualització.
Exemple	Dialogyca BDDH (Univ. Complutense).

1.4. Sistemes de gestió bibliogràfica o gestors bibliogràfics

Els acadèmics i els investigadors

Diferenciem entre acadèmics i investigadors perquè ni tots els acadèmics són investigadors ni tots els investigadors són acadèmics. Per exemple, els estudiants dels últims cicles d'universitat sens dubte són acadèmics, però els seus treballs no són necessàriament fruit de la investigació, sinó més aviat d'estudis avançats. D'altra banda, hi ha molts investigadors d'empreses i corporacions que no són a l'acadèmia, ja que no treballen a la universitat ni imparteixen docència.

Es tracta d'una classe de sistemes de gestió documental centrada en una necessitat molt característica (i possiblement exclusiva) del col·lectiu acadèmic i investigador: l'emmagatzematge de referències bibliogràfiques, la seva posterior recuperació de manera selectiva i la generació de bibliografies amb diferents formats.

Com ja se sap, una de les característiques del treball acadèmic és la necessitat de publicar. És el famós *publish or perish* (o publiques o mors). Els membres de l'acadèmia, siguin estudiants de cicles superiors, professors d'universitat o només investigadors, poden justificar la seva carrera acadèmica i els seus avanços en la investigació a través de la publicació d'articles en revistes científiques.

El punt important aquí consisteix en el fet que aquests articles s'han de basar en coneixements anteriors. La ciència es defineix com a coneixement acumulatiu. És més, un dels criteris que es manegen per separar la ciència de la pseudociència és aquest: si algú afirma que els seus descobriments no deuen res als coneixements anteriors, segur que estem davant un estafador.

La qüestió és que a cap científic ni acadèmic seriós se li acut treballar sobre el buit. A l'inrevés, part de l'èxit d'una investigació en concret, o de tota una vida acadèmica en general, descansa en l'habilitat de l'investigador per analitzar la producció científica anterior en el seu camp.

Això ens condueix a la necessitat clàssica d'investigadors i acadèmics de manejar citacions i bibliografies. Aquest col·lectiu és molt nombrós, almenys a nivell internacional, i aquesta necessitat és imperiosa. Probablement, aquests dos factors expliquen que els sistemes de gestió bibliogràfica es trobin entre els més populars en el món de la gestió documental (i de les aplicacions de la Web 2.0).

1.4.1. Estructura i característiques

Un sistema de gestió bibliogràfica sol tenir aquests quatre grups funcionals:

1) Un conjunt de mitjans per a l'entrada de referències bibliogràfiques

Això inclou l'entrada manual a través de registres predefinits per als tipus documentals habituals en el món acadèmic: articles de revista, llibres, ponències i comunicacions en congressos, capítols de llibre, etc. Però el més important és que cada vegada més els sistemes de gestió bibliogràfica poden importar referències de manera automàtica des de diferents fonts d'informació, típicament bases de dades, catàlegs, cercadors, etc.

2) Un sistema de cerca

Cap sistema documental no està complet sense un sistema de cerca que permeti crear grups selectius de referències. Aquesta funció queda assumida per la cerca simple o avançada del sistema que permetrà a l'usuari crear pàgines de resultats amb referències bibliogràfiques basades en paraules clau, en noms d'autors, en títols de revistes, etc.

3) Un sistema per a generar bibliografies

A diferència d'altres sistemes documentals, l'objectiu final no és solament la recuperació d'informació, tot i la seva importància cabdal, sinó generar una bibliografia configurada d'acord amb una norma o un format específic.

Exemple

El format de bibliografia que ens exigeix la revista en què pretenem publicar l'article o la que marca la universitat on volem defensar un treball acadèmic.

La bibliografia com a resultat final, al seu torn, pot ser generada en format imprès o com a arxiu informàtic. En aquest últim cas, sol existir la possibilitat de triar entre diverses formes de codificació: arxiu ascii, doc (Word), odt, rtf, html, xml, etc.

4) Finalment, un sistema per a introduir cites i posteriorment generar la bibliografia de manera automàtica

Els sistemes d'última generació solen proporcionar un *plug-in* o connector per a editors de text (Word o OpenOffice, típicament). Gràcies a aquest connector, és possible incorporar cites en el text d'un treball acadèmic (article, tesi, etc.) mitjançant cerques en la base de dades fetes des de l'editor de textos. Posteriorment, el sistema genera la bibliografia completa, formatada i ordenada de manera adequada, al final del document.

1.4.2. Mercat i exemples: sistemes d'escriptori contra sistemes en línia

Els programes més importants d'aquesta categoria es divideixen en:

1) Sistemes d'escriptori, és a dir, programari convencional que resideix en el disc dur de l'usuari i es carrega des d'aquest i

2) Sistemes en línia, és a dir, programari que no és necessari instal·lar en l'ordinador de l'usuari, sinó que s'executa a través d'un navegador d'Internet.

Al final, tots dos s'acaben executant en la memòria RAM de l'ordinador de l'usuari, però a efectes pràctics hi ha grans diferències entre els programes d'escriptori i els programes en línia.

Possiblement, les aplicacions en línia estan cridades a canviar (a revolucionar?) l'ofimàtica dels propers anys. Dit d'una altra manera: probablement, la futura versió d'Office (MS o OpenOffice) ja no s'executarà en mode local, sinó que ens connectarem a un servidor web i l'executarem des d'un navegador; després, naturalment, d'identificar-nos i entrar en el nostre espai personal. Només ocasionalment farem ús de la versió d'escriptori corresponent, just al contrari que ara.

De fet, ja és la manera més habitual de treballar per als sistemes que ens ocupen ara, els gestors d'informació bibliogràfica, però també per als gestors de continguts i cada vegada més per als gestors i editors d'imatges.

La taula següent és un resum de característiques de tots dos tipus d'aplicacions (en línia i d'escriptori).

Taula 5. Característiques dels gestors bibliogràfics

Característica	Escriptori	En línia
Manteniment de l'aplicació	A càrrec de l'usuari final	A càrrec del proveïdor del servei
Ús de l'aplicació	Mode local des d'un ordinador amb una aplicació necessàriament preinstal·lada i configurada	Qualsevol ordinador de qualsevol lloc del món sense necessitat d'instal·lació prèvia de cap aplicació
Ubicació de les dades	Disc dur d'un ordinador concret accessible només en mode local	Disc dur d'un servidor web accessible des de qualsevol navegador i per diversos usuaris. Disc dur de l'usuari si ho desitja
Seguretat	Típica d'usuaris finals (feble i contradictòria)	Típica d'organitzacions professionals (forta i sistemàtica)
Velocitat potencial	Molt alta (limitada només pel maquinari de l'usuari)	Alta/mitjana/baixa (limitada pel tipus de connexió a Internet)
Funcions	Sense límits *a priori*	Limitades per problemes logístics

Aplicacions en línia

Dins dels gestors bibliogràfics en línia, dues de les aplicacions més importants, per la seva presència en universitats i centres d'investigació, són RefWorks (http://www.refworks.com) i EndNote Web (http://www.endnoteweb.com). En aquest cas, es tracta d'aplicacions comercials, és a dir, la seva utilització requereix una subscripció prèvia per part de la institució (universitat o centre d'investigació), que al seu torn queda a la disposició dels investigadors o usuaris individuals de la institució.

En el cas de RefWorks i EndNote es tracta de sistemes funcionalment molt similars i que representen probablement el nivell més alt de l'estat de la qüestió en aquesta tecnologia per l'excel·lència de les seves funcions i prestacions. Les seves úniques diferències procedeixen de les empreses respectives, totes dues lligades a la indústria de les bases de dades.

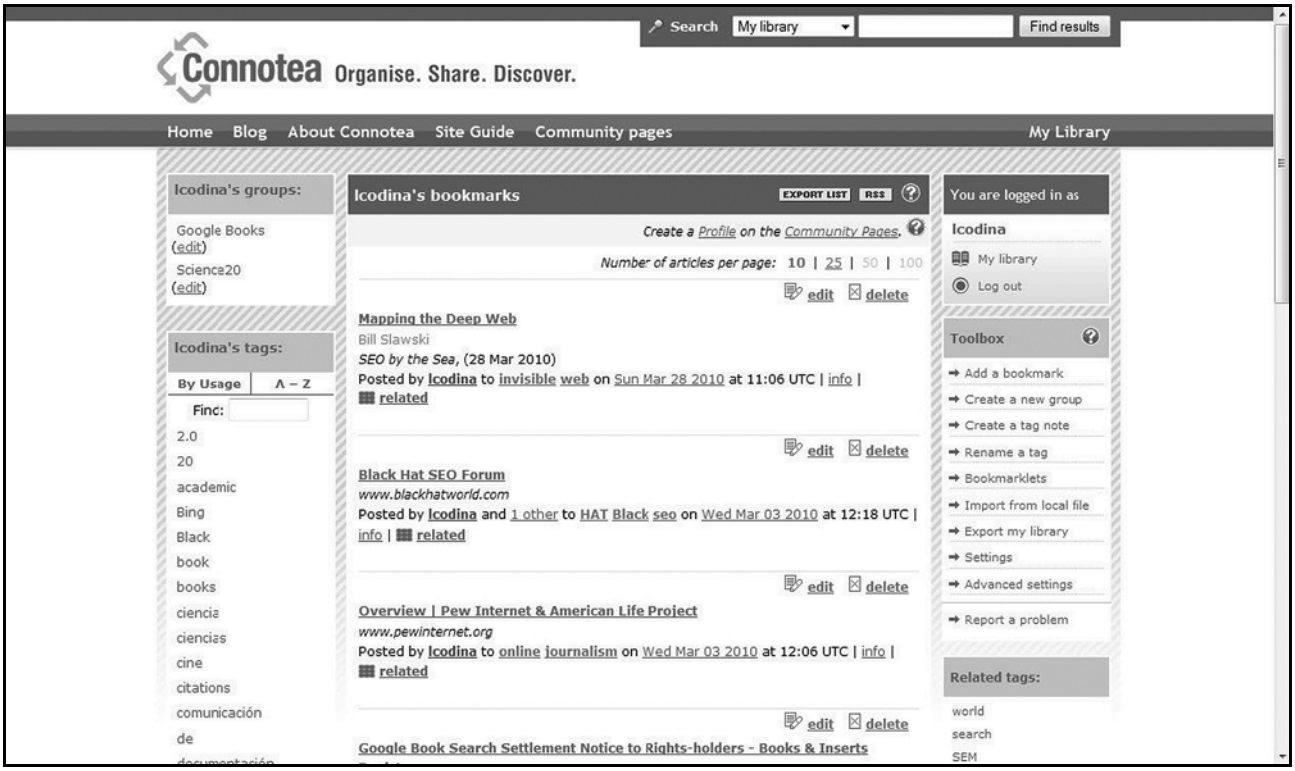
RefWorks és un producte del productor i distribuïdor de bases de dades acadèmiques ProQuest, mentre que EndNote Web és un producte de Thomson Reuter. Totes dues són empreses productores i distribuïdores de les bases de dades científiques més importants del món.

A més de les dues anteriors, hi ha un grup d'aplicacions en línia per a la gestió bibliogràfica de cost zero que presenten prestacions molt notables, en alguns casos comparables a productes comercials com els assenyalats abans. Els més importants són Zotero (www.zotero.org) –en realitat és una aplicació que combina un programari d'escriptori i un sistema en línia–, Connotea (www.connotea.org), CiteULike (www.citeulike.org) i Mendeley (www.mendeley.com).

La característiques comunes a aquests quatre programes gratuïts són les següents:

- Extremada facilitat d'ús.
- Plena integració amb el navegador d'Internet.
- Extrema facilitat per importar informació de pàgines web.

Figura 4. Interfície del sistema de gestió bibliogràfica Connotea



Aplicacions d'escriptori

Es dóna la circumstància que totes les aplicacions de gestió bibliogràfica d'escriptori més importants pertanyen a la mateixa empresa (per un procés d'adquisicions al llarg dels últims anys), en concret a una de les divisions de programari de Thomson Reuters, i es tracta de les següents:

- Reference Manager (www.refman.com)
- ProCite (www.procite.com)
- EndNote (www.endnote.com)

El motiu pel qual una mateixa empresa produeix i manté tres productes similars és un misteri per a aquests autors. Tot el que es pot dir és que hi ha una lleugera diferenciació de l'usuari final de cada producte. Reference Manager és el sistema més complet i més complex i per tant està destinat a un usuari expert, mentre que ProCite és probablement el sistema més fàcil d'utilitzar a costa de prescindir d'algunes prestacions i per tant està destinat a usuaris que no desitgen aprofundir en l'ús d'una aplicació. Finalment, EndNote se situa en algun punt intermedi i és, alhora, el que disposa d'una versió en línia, com hem vist abans.

1.5. Sistemes d'indexació

Els motors de cerca, també denominats *indexadors* o *sistemes d'indexació*, s'han fet justament famosos arran de l'important paper que exerceixen els cercadors d'Internet, especialment *Google*. Aquests serveis, que faciliten l'accés al text

complet dels documents que es troben a Internet, disposen d'un motor de cerca (*search engine)* que facilita la consulta de qualsevol terme o combinació de termes que apareguin en les pàgines web i altres documents (pdf, per exemple) que troba a Internet.

La veritat és que en l'inici de la web, aquests sistemes d'indexació ja existien. Els seus antecedents es remunten a les primeres bases de dades de text complet. Lexis va ser un dels primers sistemes que va oferir accés al text complet dels documents que contenia. Això passava entre finals dels anys setanta i a principis dels vuitanta. És coneguda l'existència d'altres programes d'aquest estil que funcionaven amb grans sistemes almenys des de la dècada dels vuitanta com STAIRS, Basis, DOCU-MASTER, etc. També durant els anys vuitanta van aparèixer les primeres versions d'aquesta classe de programes per a microordinadors: AskSam, Personal Librarian o ZyIndex en són alguns exemples.

Així doncs, els motors de cerca són un tipus d'SGD que serveix per a crear bases de dades de text complet, que elaboren uns índexs voluminosos que permeten recuperar la informació a partir de qualsevol paraula que formi part dels documents de la base de dades.

No existeix una denominació consolidada per referir-se a aquest tipus de programa. En anglès s'utilitzen els termes *text retrieval software* (programari de recuperació de text), *full-text retrieval system* (sistema de recuperació de text complet) o *text information management system* (sistema de gestió d'informació textual), entre altres, per referir-se a aquest tipus d'SGD. En francès utilitzen l'expressió *moteurs d'indexation et de recherche* (motors d'indexació i de cerca).

La diferència essencial entre un sistema de gestió de bases de dades documentals (SGBDD) i un motor de cerca (o sistema d'indexació) és que en aquest últim no hi ha cap mòdul per a dissenyar i administrar models de registre. De fet, els motors de cerca no utilitzen registres en el sentit de representacions dels documents (documents secundaris), sinó que generen índexs directament a partir de l'anàlisi dels documents originals, a l'estil dels programes que hi ha darrere de Google.

1.5.1. Estructura i característiques

Per descriure el funcionament d'aquest tipus de programa partirem dels mòduls bàsics que s'han descrit anteriorment per als SGBDD (vegeu 1.3.1.), és a dir, administració col·lecció (o base de dades), manteniment, indexació, recuperació, i sortida i intercanvi d'informació.

1) Administració de la col·lecció

Els documents que formen part de la col·lecció o del fons documental es mantenen en la màquina original (ja sigui un ordinador local o un de remot). El progra-

ma d'indexació genera uns índexs a partir dels quals es pot accedir als documents de manera selectiva a partir del contingut del text complet de la col·lecció.

D'aquesta manera, la col·lecció està formada per dos tipus de dades. D'una banda, els fitxers amb els documents i, de l'altra, els índexs que remeten a aquests documents. Els documents poden estar localitzats en diverses unitats d'emmagatzematge o en servidors externs, i l'única cosa que cal tenir en compte és la seva ubicació precisa en el moment de definir la col·lecció (en quines unitats de disc o quines són les adreces dels servidors remots en els quals es troben els fitxers amb els documents) que cal indexar. Quan executem el programa utilitzarem els índexs i, amb l'apuntador del document, podrem visualitzar-los a través de l'aplicació original amb la qual van ser creats.

Encara que aquesta classe d'aplicacions no acostuma a estructurar els documents, és cada vegada més freqüent l'ús de camps o d'etiquetes que permeten donar una aparença d'estructura de camps a la col·lecció i faciliten l'accés a parts concretes del document, habitualment el títol, la data de creació o l'autor. Aquesta estructuració la poden dur a terme, tot i que no utilitzen una autèntica estructura de registres, per derivació de les metadades que solen generar cada vegada més aplicacions.

Exemple

Si el document recuperat és una pàgina web, podrem veure'l en un navegador, però, si es tracta d'un document de text, podrem veure'l en un editor de text Word o en Open Office, etc.

2) Manteniment (entrada de dades)

Tal com es dedueix del que s'ha descrit en l'apartat anterior, l'entrada de dades al sistema no s'acostuma a fer des del teclat perquè ja es disposa dels fitxers informàtics amb la informació que s'ha de processar (documents html, documents de text, fulls de càlcul, gràfics, etc.).

El problema pot provenir de la diversitat de formats en els quals poden estar els documents que han de formar part de la base de dades (o col·lecció), que poden ser de tot tipus (doc, odt, rtf, html, xlc, pdf, eds, tiff, etc.).

3) Indexació

El programa indexa el text complet dels documents que formen part de la base de dades o col·lecció i també, si n'hi hagués, els indicadors, marques de camp o metadades. D'aquesta manera es poden delimitar les consultes a un camp determinat del registre.

4) Recuperació

En general, el procés de consulta en sistemes d'indexació es duu a terme de manera similar a la consulta de bases de dades de tipus referencial, és a dir, s'usa l'àlgebra booleana i es disposa d'una sèrie d'operadors complementaris

(truncament, proximitat, etc.). Ara bé, a més d'aquest tipus de consulta, que és la tradicional en tots els programes de recuperació de la informació, els motors de cerca estan experimentant amb altres tipus de prestacions, fonamentalment, les cerques semàntiques i les cerques per patrons.

a) Cerca semàntica

Es tracta de poder ampliar la consulta d'un terme a tots aquells que hi estiguin relacionats d'alguna manera –derivació morfològica, equivalència lingüística, sinonímia, antonímia, generalització, especialització, etc.

Exemple

Si un usuari busca informació sobre "copyright", el sistema li mostra un conjunt de termes relacionats amb el que s'ha sol·licitat (propietat intel·lectual, drets d'autor, drets d'explotació, etc.).

b) Cerca per patrons

Es tracta d'un sistema totalment oposat a l'anterior. En aquest cas ens estem referint a una anàlisi que no té en compte la morfologia (la forma de les paraules), ni la sintaxi (l'ordre i la coordinació de les paraules) ni la semàntica (els significats), sinó que la indexació de la informació es basa en patrons de bits. D'aquesta manera, qualsevol tipus d'informació, ja sigui text, so o imatge, està indexada i es recupera emprant el mateix sistema de representació. Aquesta tecnologia es basa en l'aparença física dels termes (el seu codi binari) i no en la semàntica (el seu significat).

Les cerques per patrons permeten comparar textos o imatges a partir de patrons binaris, és a dir, permeten trobar textos o imatges que comparteixen una sèrie de característiques estructurals comunes.

Exemple

Si busquem "Eltsin", el programa ens facilitarà tots els documents en els quals aparegui exactament aquesta paraula i també aquells en els quals constin altres paraules semblants: "Yeltsin", "Elsin", "Ieltsin", etc. Podria passar el mateix amb "Gadafi", "Kadhafi", "Kadafi", "Gadaffi", etc. Les variacions es poden deure al fet que els termes hagin estat mal escrits, mal reconeguts per un OCR, o al fet que es tracti de transliteracions fetes amb criteris diferents.

5) Ponderació de resultats

La utilització sense control de les cerques per patrons i semàntiques comporta l'aparició d'un bon nombre de resultats no desitjats. Els mecanismes de ponderació de termes constitueixen un instrument complementari molt útil i pràcticament imprescindible per a minimitzar els efectes no desitjats d'aquest tipus de consultes. Cal tenir present que, en aquests entorns, normalment es recupera un nombre molt alt de documents i que cal tenir instruments que ajudin a determinar quins són els més rellevants.

1.5.2. Tipus d'aplicacions

1) Cercadors de pàgines web

Com ja hem apuntat en altres apartats, els cercadors de pàgines web han estat els que han popularitzat els sistemes d'indexació i els han donat a conèixer al gran públic. Aquests poden ser generals o a escala de tot el Web (Google, Yahoo, Bing, etc.), o particulars d'una sola seu web, com pot ser el cas del cercador del lloc de la Universitat Oberta de Catalunya (www.uoc.edu) o de qualsevol altre organisme públic o privat.

2) Bases de dades de text complet

Els sistemes d'indexació també es poden utilitzar per a crear bases de dades de text complet. En aquest cas hi pot haver una certa unitat temàtica o de publicació en el contingut de la base de dades, a diferència de l'anterior, en què podem trobar una amalgama i una varietat molt dispar.

En la major part dels casos s'utilitzen metadades descriptives del contingut (autor, títol, data, etc.) i en tots els casos s'utilitza algun tipus de metadades administratives, estructurals o de drets de propietat. Podem trobar exemples diversos en premsa, revistes científiques i fons editorials.

Exemples

Premsa:

- *MyNews* <http://www.mynewsonline.com>
- *El País* (http://www.elpais.com/archivo/)
- *La Vanguardia* (http://www.lavanguardia.es/hemeroteca/)

Revistes acadèmiques:

- *Ariadne* (http://www.ariadne.ac.uk/search/)
- *Information Research* <http://informationr.net/ir/search.html>

Fons editorials:

- Ocenet (http://consulta.oceano.com). Fons de l'editorial Océano, accessible des de la xarxa de biblioteques de la Diputació de Barcelona.
- V-lex (http://vlex.com/). Base de dades de legislació, que conté més de 30 milions de documents de 130 països.

1.5.3. Mercat

Encara que la major part d'aplicacions requereix l'estructura client-servidor, també és possible trobar algunes versions personals que funcionen amb un microordinador. A continuació es descriuen els principals programes d'aquest estil que es troben presents en el mercat espanyol.

Nom	askSam
Productor	askSam (http://www.asksam.com)

Nom	Autonomy
Productor	Autonomy <http://www.autonomy.com>

Nom	Google Custom Search
Productor	Google <http://www.google.com/coop/cse/>
	Serveix per a indexar una part de la web que sigui de l'interès de l'usuari. Versió de pagament (Site Search) i versió gratuïta.

Nom	Google Search Appliances
Productor	Google <http://www.google.com/enterprise/public_search.html>
	Motor de cerca per a servidors propis. Aplicació per a Intranet o seus web. Hi ha una versió reduïda (Mini Search).

Nom	Greenstone
Productor	New Zealand Digital Library Project at the University of Waikato (www.greenstone.org). Programari lliure.
Exemple	New Zealand Digital Library <http://nzdl.sadl.uleth.ca/cgi-bin/library>

Nom	Swish-e
Productor	Comunitat de desenvolupadors (http://swish-e.org/). Programari lliure.
Exemple	Ariadne (http://www.ariadne.ac.uk/search/)

Nom	Apache Lucene
Productor	Apache (http://lucene.apache.org/)
Exemple	Llista a: http://wiki.apache.org/lucene-java/poweredby

2. Distribució de bases de dades

La producció i la distribució de bases de dades són dos processos complementaris que de vegades executen agents diferents, amb tecnologia i eines ben diferenciades. En l'apartat anterior ens hem centrat en *la producció* de bases de dades, és a dir, en el procés de creació i elaboració d'uns continguts informatius que queden estructurats d'una manera determinada i que són explotables amb el concurs d'un sistema informàtic. La *distribució*, en canvi, és el conjunt d'operacions que facilita als usuaris l'accés a aquests continguts informatius. Així doncs, mentre que el procés de producció permet elaborar un contingut únic, el procés de distribució permet que pugui arribar al seu públic per diferents canals.

Aquestes diferències essencials entre tots dos processos expliquen que les estratègies i els programes informàtics relacionats amb la producció, normalment tenen poc a veure amb els mecanismes i instruments que s'utilitzen per a la distribució.

Fins fa pocs anys, els productors i els distribuïdors de bases de dades (aquests últims en particular) acostumaven a tenir un caràcter especialitzat i a tenir, per tant, una estructura empresarial i professional que els recolzava. Aquesta situació ha canviat radicalment amb l'eclosió d'Internet i el desenvolupament de diferents eines fàcilment configurables i adaptables que posen a l'abast de centres d'informació i documentació petits i mitjans, i fins i tot d'usuaris personals, la possibilitat de convertir-se en productors i distribuïdors de bases de dades. Aquestes eren unes funcions que, amb la tecnologia anterior, eren molt difícils d'exercir sense una infraestructura molt especialitzada i costosa. En l'exposició que fem aquí tindrem en compte aquest canvi important.

2.1. Antecedents

La consulta local va ser el primer sistema que es va utilitzar per a facilitar l'accés dels usuaris a la base de dades. Té limitacions importants, fonamentalment d'accés, ja que obliga els usuaris a desplaçar-se expressament al centre de documentació o a la unitat on l'aplicació està disponible.

De manera complementària, especialment en centres de l'àmbit de les humanitats i les ciències socials, també es va utilitzar la publicació de bibliografies impreses com a forma per a la distribució dels continguts de bases de dades. Aquestes bibliografies es poden elaborar automàticament des d'alguns SGD i consten, bàsicament, de dues parts: en primer lloc, una llista global correlativa dels registres numerats o ordenats per algun element descriptiu, normalment l'autor, i que inclou la descripció completa de cadascun dels registres; en segon lloc, es poden trobar índexs diversos –autors, títols, matèries, etc.– que remeten al número de

registre de llista general. Els inconvenients més destacables són els alts costos d'impressió i de distribució i les dificultats per actualitzar les obres.

Finalment, una altra via que s'ha utilitzat en el passat és el suport òptic, que implicava incorporar al disc compacte el programa de recuperació de la informació (o, almenys, el mòdul de consulta). Aquesta via va ser notablement usada a la fi del s. XX i després va passar ràpidament a la història amb l'eclosió del web.

Bibliografies impreses

Inmagic, ISIS, Pro-Cite i, en menor mesura, Knosys, tenen sistemes per a facilitar, més o menys, aquesta tasca d'accés a la base de dades.

2.2. Web

El web és el sistema de distribució de bases de dades documentals més utilitzat. El motiu és senzill: l'usuari que consulta la base de dades només ha de tenir un navegador per poder accedir als registres de manera actualitzada i disposar, en alguns casos, de les mateixes prestacions de consulta i explotació que tenen els sistemes de gestió documental.

És a dir, l'usuari no necessita instal·lar cap versió client del programa que gestiona la base de dades, sinó que és el mateix navegador d'Internet (Internet Explorer, Mozilla o Opera) el que actua com a client de la base de dades. Des del navegador, tan sols haurà d'indicar la seva petició mitjançant un formulari html per a rebre les respostes també en aquest format que el navegador no tindrà cap dificultat a reproduir en el monitor de l'usuari.

Formulari HTML

És una secció d'una pàgina web que conté elements de control (caselles de verificació, botons d'elecció, etc.) i permet introduir text perquè un servidor web el processi.

Ara bé, perquè aquest mètode d'accés sigui possible, és necessari, en el costat del servidor, un programa o un conjunt de programes que permetin establir la comunicació entre dos entorns en principi incompatibles o diferents: la base de dades gestionada pel SGBD, d'una banda, i el navegador web que utilitza l'usuari i que només és capaç d'interpretar pàgines html transmeses mitjançant el protocol http, de l'altra. Aquests programes solen rebre la denominació CGI (*common gateway interface*) o interfície de passarel·la.

Terme *passarel·la*

S'utilitza el terme passarel·la (*gateway*) perquè fa referència a la funció de relació entre el servidor web i les aplicacions externes.

2.2.1. Estructura

Els elements bàsics que intervenen en aquest procés son els següents:

- Navegador (per exemple, Explorer, Mozilla, etc.).
- Servidor httpd (per exemple, Apache, Internet Information Server, etc.).
- Programa CGI (per exemple, WWWIsis, WebPublisher, etc.).
- Interfície de consulta.
- Base de dades.

Els programes (el servidor httpd i el CGI) estaran instal·lats en un servidor, que tindrà targeta de xarxa i una adreça IP.

De tots els elements enumerats, potser el menys conegut és el programa CGI, que actua com a sistema de comunicació o passarel·la entre els registres de la base de dades, que no estan codificats en html, i el navegador web, que només pot interpretar informació codificada en html. El CGI és un protocol estàndard desenvolupat originalment per a Unix. La creació d'aquesta especificació va ser obra dels autors principals dels servidors http (Tony Saunders, entre altres) i s'explica perquè no volien haver d'anar ampliant constantment les funcions dels servidors per anar adaptant-los als programes nous. Per això van preferir crear un nucli per al servidor web i proporcionar-li un instrument que li permetés estendre els seus serveis i les seves capacitats.

Així doncs, el protocol CGI és un estàndard per mitjà del qual un servidor web (httpd) es pot comunicar amb un programa extern i s'obtenen documents html dinàmics (és a dir, que es generen al moment, ja que varien segons quina hagi estat la petició de l'usuari). Aquest protocol estableix una manera d'enviar dades des d'una pàgina web –per mitjà d'un formulari– i de processar-les mitjançant un fitxer executable –programa CGI– que està situat en el directori cgi-bin, o equivalent, d'un servidor.

D'altra banda, un programa CGI és una aplicació informàtica escrita en llenguatge de programació (Perl, C, C++, etc.) que posteriorment és executada i interpretada per un servidor web per poder contestar peticions d'informació dels usuaris. El programa CGI és capaç de llegir i interpretar les ordres que se li transmeten des d'un formulari html, algunes introduïdes per l'usuari (p. ex. els termes de cerca) i altres corresponents a paràmetres generals (p. ex. la ubicació del programa i de la base de dades en el servidor, el format de visualització, el nombre de documents a visualitzar, etc.). A continuació, els executa i transfereix el resultat a l'usuari en format html.

A més del programa CGI és necessari preparar una interfície de consulta adaptada a la base de dades que tingui en compte els camps que s'han definit, els formats de visualització, etc. Aquesta interfície es construeix amb el llenguatge de programació del programa CGI, entremesclat amb codi html, i consta bàsicament de tres elements: pantalla de consulta, pantalla de visualització de resultats (llista) i pantalla de visualització del document complet.

Sobre aquesta interfície, vegeu l'apartat 2.3 de aquest mòdul didàctic.

2.2.2. Mercat

Els programes CGI que s'indiquen a continuació serveixen per a poder consultar bases de dades documentals que han estat creades amb el SGBD corresponent (FileMaker, Knosys, Inmagic o CDS/ISIS).

Nom	FileMaker
Productor	Claris <www.filemaker.com>
Distribuïdor	Claris <www.filemaker.com/es>

Comentaris	• Té un assistent que permet elaborar ràpidament la interfície de consulta.
Exemples	BIGPI (Geologia de la península Ibèrica): http://www.bib.ub.edu/fileadmin/bigpi/bigpi.htm

Nom	Knosys Internet
Productor	Micronet <http://www.micronet.es/menu/prof/mki.htm>
Distribuïdor	Micronet
Comentaris	• Té un assistent, encara que és una mica limitat. • No permet ordenar els registres per cap criteri.
Exemples	En l'apartat "Clients" de les pàgines dedicades a Knosys Internet es pot trobar una llista d'usuaris.

Nom	WebPublisher
Productor	Inmagic <http://www.inmagic.com>
Distribuïdor	Doc 6 <http://www.doc6.es>
Comentaris	• Té un bon assistent. • Es poden mostrar els índexs de camp.
Exemples	Directori de bases de dades: <http://www.andornot.com/webpublinks>

Nom	WWWIsis + GenIsis
Productor	Bireme <http://regional.bvsalud.org>
Distribuïdor	Bireme
Comentaris	• Té assistent (GenIsis). • Es poden mostrar els índexs de camp.
Exemples	Bases de dades de BVS: <http://bases.bvs.br>

2.3. Interfície de consulta

La interfície de consulta d'una base de dades serveix per a establir la comunicació entre persones que busquen informació i els sistemes de recuperació de la informació i és una de les parts més importants del procés de distribució d'una base de dades. Com ja hem avançat, en el cas de base de dades distribuïdes a través del Web, aquesta interfície està formada per un conjunt de pàgines HTML de les quals podríem destacar les cinc classes següents:

- El format de consulta.
- Els resultats.
- La visualització del document complet.
- La informació general.
- Les ajudes.

En el cas de les interfícies de consulta de bases de dades web, destaquem dos textos que han abordat el procés de consulta a bases de dades des del punt de vista del procés seguit per l'usuari (Marchionini, 1995; Shneiderman, 1997) i

també els dos textos clàssics en l'àmbit de la usabilitat i l'arquitectura de la informació (Nielsen, 2000, 2006; Morville i Rosenfeld, 2006). Aquests últims disposen d'apartats dedicats a la interfície de consulta de bases de dades en els quals es destaquen els aspectes principals que ha de complir una bona pàgina d'aquest tipus. A partir d'aquests precedents, es va presentar una proposta d'indicadors sobre bases de dades (Abadal, 2002).

2.3.1. Què és una interfície de consulta

Una interfície de consulta és un conjunt d'elements de programari i de maquinari que serveix per a establir la comunicació entre persones que busquen informació i un o més sistemes de recuperació d'informació.

Hi ha unes altres dues definicions complementàries que expressen de manera més precisa aquesta aproximació. En la primera, Marchionini es refereix a la interfície des d'un punt de vista conceptual i la vincula amb el procés de cerca, en general:

> "La interfície ha de proporcionar un mapatge (*mapping*) robust entre el contingut de la base de dades i les representacions conceptuals que el cercador d'informació manipula." (Marchionini, 199, pàg. 39)

Per tant, Marchionini posa l'èmfasi en el fet que la interfície serveix per a establir la comunicació entre persones amb necessitats d'informació i un sistema de recuperació d'informació.

En la segona, Marti Hearst fa una aproximació més pragmàtica i precisa amb més detall quina és la funció concreta –els objectius– d'una interfície de consulta i la vincula amb les fases del procés de cerca:

Procés de cerca

Comprensió (definició del problema), planificació (selecció d'un sistema de cerca, formulació d'una pregunta, execució de la cerca), avaluació i ús.

> "La interfície d'usuari ha d'ajudar a comprendre i expressar les necessitats d'informació. També ha d'ajudar l'usuari a formular les seves preguntes, seleccionar entre les fonts d'informació disponibles, entendre els resultats i seguir el progrés de la seva cerca." (Hearst, 1999, pàg. 257)

Com ja s'ha descrit anteriorment, el nostre objectiu consisteix a determinar quins són els elements que han de formar part de la interfície de consulta.

En el nostre cas partirem de les classes de pàgina que s'han d'elaborar perquè l'aplicació funcioni de manera adequada com a interfície de consulta. Els agruparem de la manera següent:

- Consulta.
- Llista de resultats.
- Visualització del document complet.
- Altres pàgines (informació general, ajuda, etc.).

Així doncs, determinarem quins són els elements bàsics que han d'estar presents en cadascuna de les cinc classes de pàgines esmentades abans per

a contribuir a facilitar el procés de recuperació de la informació per part dels usuaris.

Taula 6. Elements d'una interfície de consulta a bases de dades

Pàgina	Components
Consulta	• Identificació de la pàgina o base de dades. • Nivells: simple, avançada, índexs. • Especificació de la base de dades (o del fons o col·lecció o subseu web). • Sistema de recollida d'informació de l'usuari. • Delimitació de la cerca a un camp o conjunt de camps. • Utilització dels operadors (booleans i d'altres). • Visualització dels índexs. • Informacions breus per ajudar en la consulta. • Elecció de la forma de presentació dels resultats: – Format de visualització. – Nombre de registres a visualitzar. • Elecció del sistema d'ordenació dels resultats. • Botons per a l'execució d'accions. • Registre de les cerques realitzades (historial). • Accés multilingüe. • Navegació entre pàgines de la interfície. • Dades identificatives (productor, data, lloc, etc.).
Llista de resultats	• Identificació de la pàgina o base de dades. • Informació sobre el terme de cerca i els resultats obtinguts. • Llista amb la descripció bàsica dels documents. – Estructura. – Inclusió del nom del camp. – Casella de selecció. • Indicació del tipus de document (objecte). • Agrupament dels resultats per categories. • Forma de presentació dels resultats: – Format de visualització. – Nombre de registres a visualitzar. • Informació sobre errors o absència de resultats. • Opcions de gestió dels registres o documents. • Elecció del sistema d'ordenació dels resultats. • Reformulació de la cerca. • Possibilitat de trobar documents similars. • Navegació entre registres de la base de dades. • Avanç i retrocés en les pàgines de resultats. • Navegació entre pàgines de la interfície.
Visualització dels documents o registres	• Identificació de la pàgina o base de dades. • Indicació del número de registre que s'està visualitzant. • Opció de canvi de format de visualització. • Ressaltar els termes de cerca. • Diferents resolucions. • Navegació entre registres de la base de dades. • Avanç i retrocés entre els registres seleccionats. • Navegació entre pàgines de la interfície.

2.3.2. Consulta

Les pàgines de consulta contenen els formularis que tenen per objectiu facilitar la recuperació de la informació continguda en la base de dades. La seva fun-

ció és permetre que l'usuari formuli la seva necessitat d'informació (una de les fases fonamentals del procés de cerca) i per això contindrà diversos quadres de text perquè es puguin introduir els termes de cerca, així com els operadors de cerca disponibles.

Ara bé, atenent la recomanació d'adaptar-se al nivell de l'usuari, ja sigui expert o principiant, seria desitjable que es poguessin tenir, almenys, dos tipus de pàgina de consulta: una consulta simple, amb poques opcions de cerca, i una consulta avançada, en la qual es puguin usar tots els operadors i, a més, combinar diversos termes. D'altra banda, també és recomanable poder consultar els índexs de camp i accedir-hi per categoria temàtica.

Figura 5. Exemple de pàgina de consulta simple (JStor)

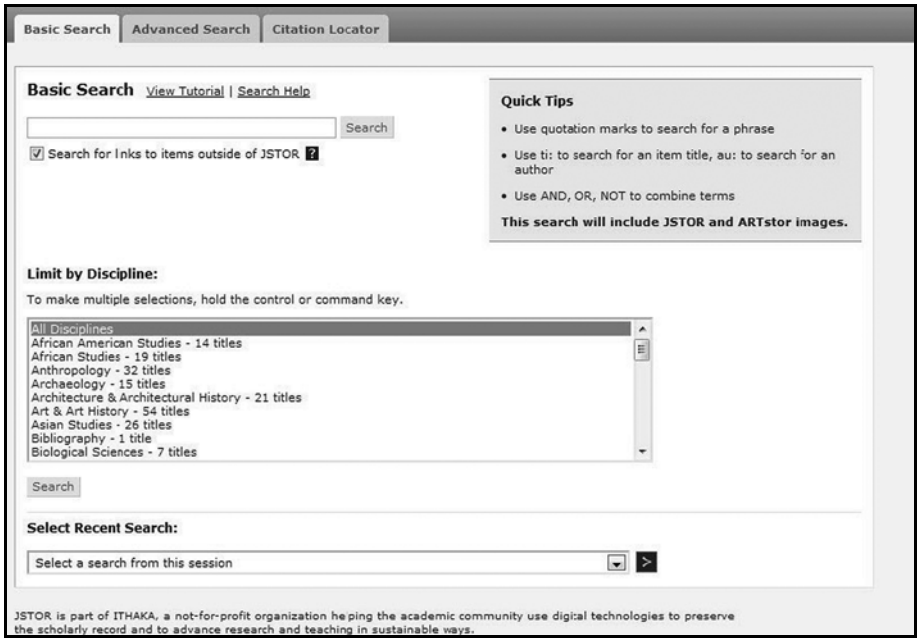


Figura 6. Exemple de pàgina de consulta avançada (JStor)

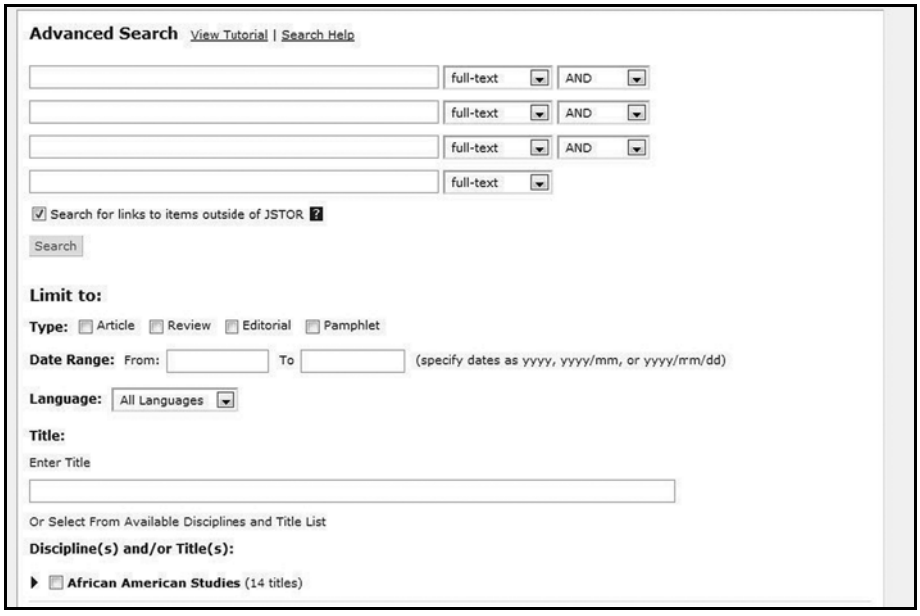


Figura 7. Exemple (vista parcial) de pàgina de consulta avançada d'un cercador acadèmic (Scirus)

Com hem detallat a la taula 6, els elements d'una pàgina de consulta poden ser bastant nombrosos i la seva composició pot generar confusió si no estan ben estructurats. Per això, és convenient establir una sèrie de zones o àrees ordenades jeràrquicament per a facilitar la seqüència d'accions que segueix un usuari quan formula una pregunta a una base de dades. En aquest sentit, cal destacar de manera especial el sistema de recollida de dades de l'usuari (formulació de la pregunta) i, en segon lloc, les especificacions de visualització, ja que l'usuari pot tenir un interès especial a escollir algunes opcions relacionades amb les característiques de visualització de la llista de resultats: quants registres es presentaran, en quin format, en quin ordre apareixeran, etc.

2.3.3. Llista de resultats

La primera resposta del sistema a una consulta expressada per l'usuari ha de ser una pàgina amb la llista que conté la informació bàsica dels documents o registres que satisfan la pregunta, és a dir, que són rellevants a la necessitat d'informació. L'objectiu d'aquesta pàgina ha de ser presentar una visió global dels resultats i facilitar a l'usuari la valoració de l'interès de cada document a partir de la seva descripció resumida.

La visualització dels resultats es pot presentar de manera textual (alfanumèrica), amb la informació dels camps que es visualitzen un darrere de l'altre o dins d'una taula per a estructurar millor l'espai de resposta. També es poden utilitzar presentacions de caràcter gràfic.

La selecció dels registres que són de més interès per a l'usuari és més complexa com més gran és el nombre de documents recuperats. En aquest punt és molt útil disposar de sistemes d'ordenació dels resultats basats en la rellevància o altres criteris lliurement configurables per l'usuari (data, autor, títol, etc.).

Figura 8. Exemple de pàgina de resultats (ISI)

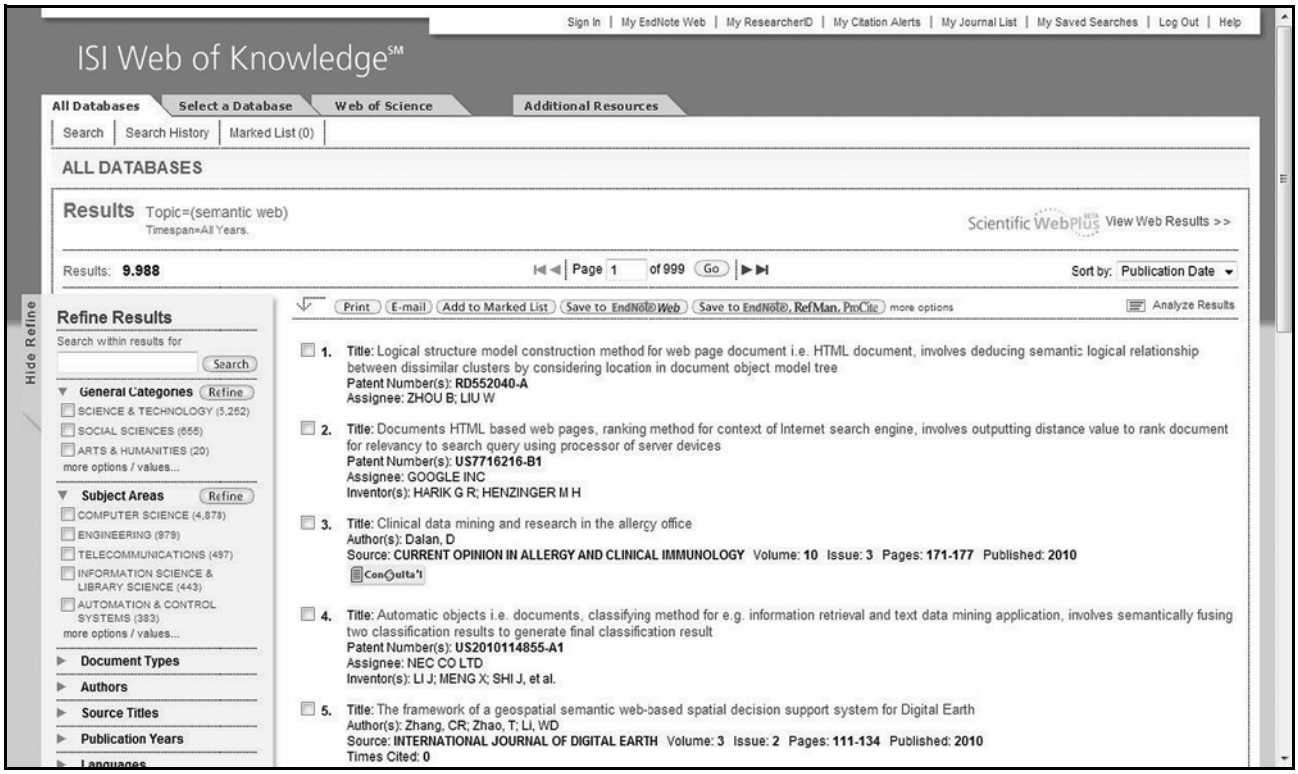
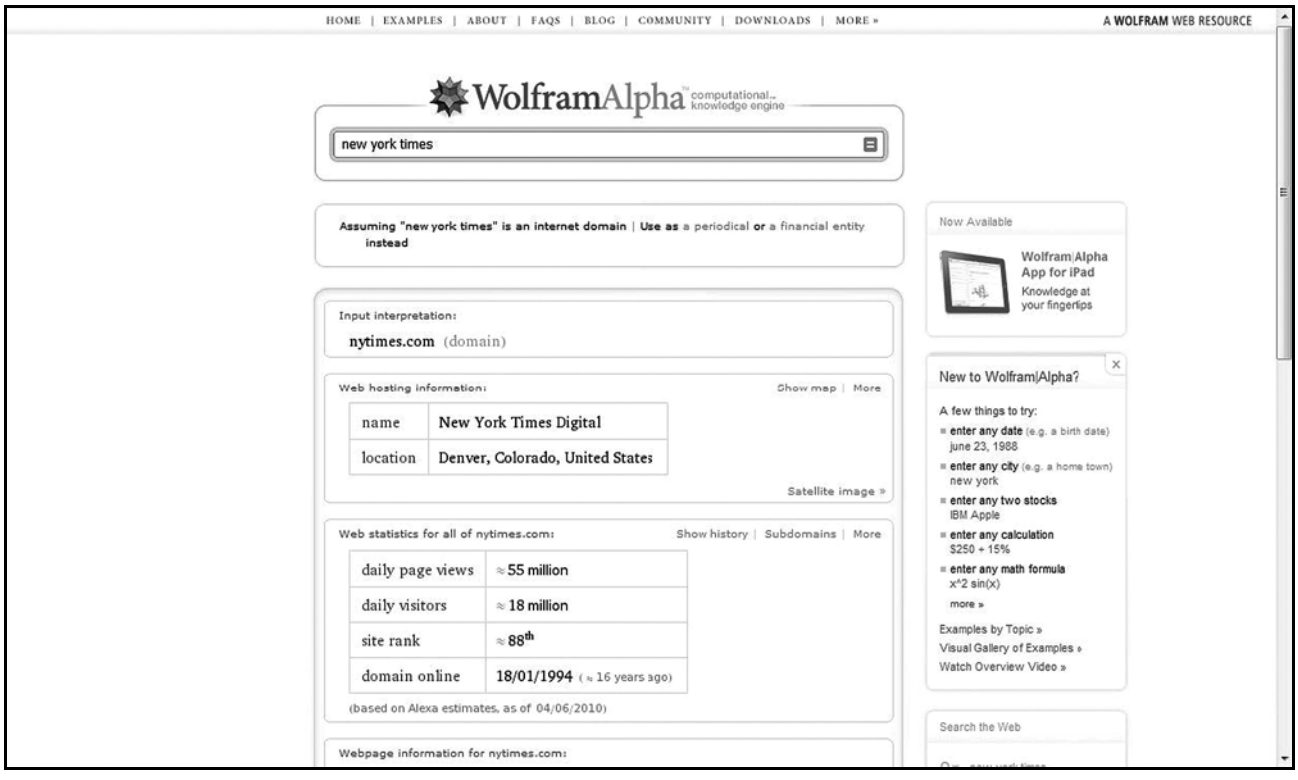


Figura 9. Exemple de pàgina de resultats de tipus tabular (Wolfram Alpha)

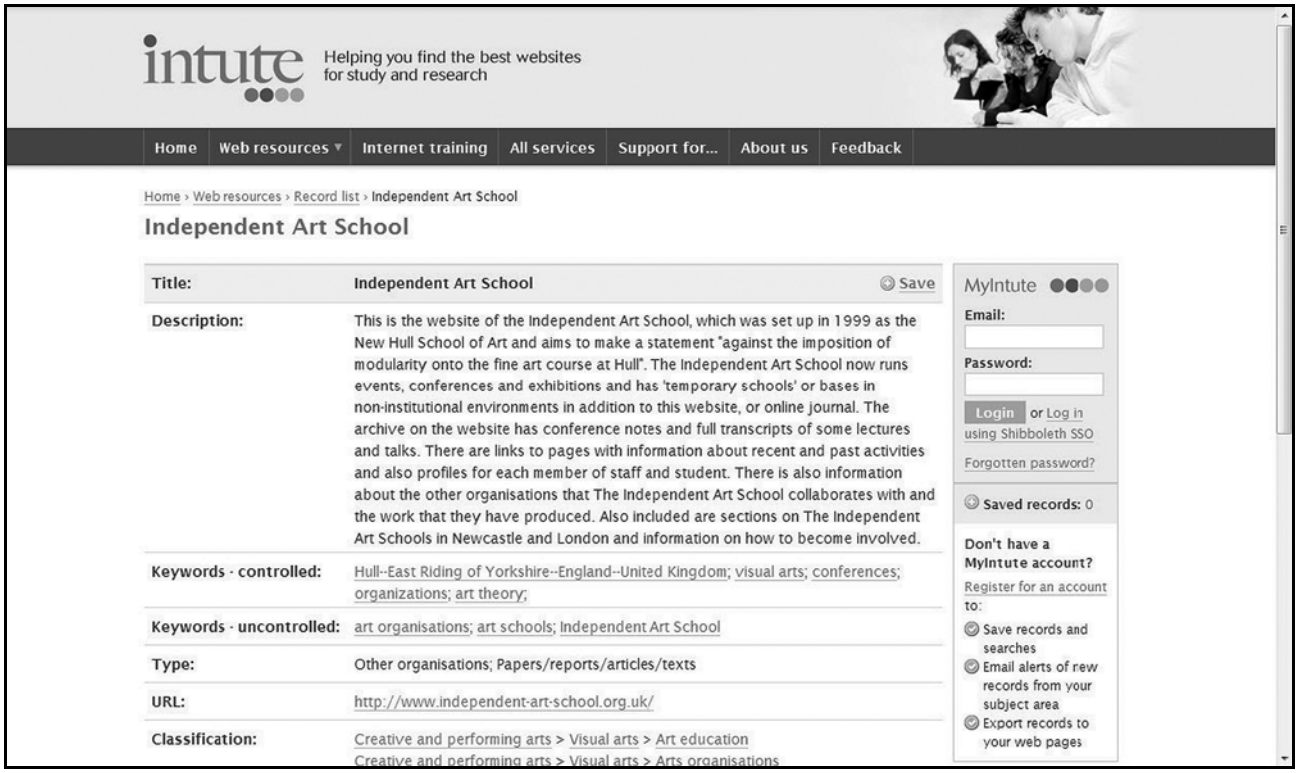


La unitat d'informació de la llista de resultats acostuma a ser el document (pàgina web, pdf. etc.) o el registre (metadades); en aquest últim cas, quan busquem en bases de dades estructurades.

2.3.4. Visualització dels documents (o registres)

Des de la pàgina de llista s'ha de passar a una altra pàgina que inclogui la visualització del document unitari sol·licitat. Aquest document pot ser de tipus textual, gràfic o sonor, o combinar algun d'aquests tipus d'informació. A més, es pot demanar la visualització de la referència o de les metadades.

Figura 10. Exemple de pàgina de document, en aquest cas un registre amb metadades (Intute)



2.3.5. Altres pàgines

El grup de pàgines descrites anteriorment constitueix el nucli fonamental de la interfície de consulta. De totes maneres, hi ha altres pàgines que les complementen i entre les quals destaquem les següents:

1) Descripció general del contingut

Aquesta pàgina informa l'usuari sobre l'àmbit geogràfic, temàtic i lingüístic de la base de dades. A més, inclou dades sobre la seva estructura (camps, etc.), el nombre de registres, etc. Les dades que proporciona han de permetre contextualitzar el contingut de la base de dades, així com mostrar-ne l'abast.

2) Ajudes

En aquest apartat s'inclouen, d'una banda, els textos que informen l'usuari sobre el funcionament de l'aplicació (és a dir, com cal fer les consultes, quines són les opcions disponibles del sistema, etc.) i, de l'altra, els missatges d'ajuda i d'error que el sistema va facilitant a l'usuari a mesura que aquest va executant les seves accions.

Nota

Com és ben sabut, les característiques fonamentals que ha de complir el sistema d'ajuda són les següents: fàcils de localitzar, ben organitzades i contextualitzades.

3) Pàgina d'identificació (connexió/desconnexió)

En algunes aplicacions és necessari incloure una pàgina que permeti a l'usuari connectar-se al sistema per mitjà d'una identificació (*login*) i una contrasenya (*password*) i, posteriorment, desconnectar-se.

La falta d'inclusió d'algun dels elements abans esmentats resta efectivitat a la interfície de consulta i dificulta a l'usuari les operacions d'accés i recuperació dels continguts de la base de dades. En qualsevol cas, cal tenir presents dues qüestions addicionals:

a) Jerarquització

És evident que no totes aquestes funcionalitats tenen la mateixa importància i que n'hi ha unes que pesen molt més que les altres. Aquesta qüestió, tanmateix, ha estat esbiaixada en la discussió que s'ha presentat.

b) Universalitat

D'altra banda, tots aquests elements tampoc no són útils ni necessaris per a tot tipus d'usuaris. Si considerem, com a mínim, dos nivells d'experiència entre els usuaris –novells i experts– es comprendrà ràpidament que, per als primers, la major part dels elements s'ha de presentar ja configurada, sense deixar-los llibertat d'elecció per personalitzar-los.

2.3.6. Tendències

A més dels elements comentats fins aquí, hem de fer referència a altres elements que encara no estan tan generalitzats com els anteriors.

1) Filtratge dels resultats per característiques

És cada vegada més freqüent tenir la possibilitat de filtrar els resultats trobats per tipus de document (llibres, articles, vídeos, etc.), data, idioma, autor, etc.

Exemples

- Cercador de notícies d'*El País*
  Disposa d'una cerca bàsica a partir de la qual es poden limitar els resultats per tipus d'article, data, etc.
- Google
  Es poden filtrar els resultats per tipus de document, data, etc.
- Catàleg UPC (http://bibliotecnica.upc.edu/search)
  Disposa de filtres per idioma, tipus de document, etiqueta (*tag*), autor, etc.

2) Presentació gràfica (visual) dels resultats

Pretén fer més fàcil, simple i intuïtiva la presentació dels resultats. Es tracta d'un àmbit en el qual fa molts anys que s'investiga, tot i que encara s'han obtingut pocs resultats.

Exemples

- Newsmap (http://marumushi.com/apps/newsmap/newsmap.cfm)
  Aplicat a notícies d'actualitat.
- Liveplasma (http://www.liveplasma.com/)
  Consulta de música i cinema.
- WebBrain (www.webbrain.com)
  Mostra gràficament les relacions de jerarquia entre continguts de pàgines.

3) Agrupacions temàtiques dels resultats

En alguns casos, els motors de cerca són capaços d'agrupar els resultats segons categories temàtiques generades automàticament.

Exemples

- Clusty (http://clusty.com/)
  A l'esquerra es presenten els grups que s'han creat automàticament amb els resultats obtinguts (abans es deia Vivisimo).
- IBoogie (www.iboogie.com)
  Presenta els grups en els quals es divideixen els resultats en funció de la coocurrència dels termes. Les jerarquies se subdivideixen.
- Moot (www.mooter.com)
  Indica les cerques relacionades. Apartat interessant pregunta-resposta.

4) Eines de descobriment: els nous OPAC de biblioteques

Els catàlegs de biblioteca estan incorporant interfícies de consulta que intenten reproduir els models de funcionament de les seus web més "reeixides", conegudes i valorades pels usuaris (Google i Amazon, per posar dos exemples). Aquestes interfícies permeten una consulta conjunta en les diverses col·leccions de la biblioteca (catàleg, repositori, col·leccions subscrites) i no només del catàleg, com era habitual fins fa poc, permeten el filtratge de resultats (temàtica, autor, tipus de document, etc.) i faciliten també la visualització de les portades. En anglès se'ls anomena *discovery tools*.

OPAC

Online public access catalog.

Exemples

- Aquabrowser
  Exemple: Queens library (http://www.queenslibrary.org/)

- Encore
  Exemple: Michigan State University (http://www2.lib.msu.edu/)

3. Metodologia per a la creació de bases de dades documentals

El principi general de creació de sistemes d'informació indica que tot projecte comença sempre per un disseny lògic i que, una vegada aprovat aquest, es procedeix al disseny físic o la implantació, en un procés que és tan interactiu com lineal, ja que la fase de disseny, per exemple, pot obligar a repensar aspectes de la fase d'anàlisi.

L'aspecte important aquí és que la metodologia ens diu clarament que el procés de creació d'una base de dades ha d'anar sempre des dels aspectes lògics cap als aspectes físics, i no a l'inrevés, com tanmateix sol succeir, ja que, en la pràctica, hi ha moltes maneres de violar aquest principi general a causa de mals hàbits de treball.

Així doncs, el procés de creació d'un sistema d'informació s'ha d'ajustar sempre al cicle de vida següent:

1) Anàlisi.
2) Disseny.
3) Implantació.

Una altra manera d'enfocar un projecte de desenvolupament és indicar que la direcció del disseny ha de procedir del que es coneix al que es desconeix, i no a l'inrevés, com succeeix quan es vol visualitzar el sistema d'informació abans de conèixer el sistema d'activitats humanes i el sistema de coneixement.

Finalment, i per la mateixa raó, la direcció del disseny ha d'anar d'allò general a allò específic i dels aspectes lògics als aspectes físics, i mai a l'inrevés, és a dir, mai no s'ha de començar a discutir o a considerar qüestions concretes (com s'imprimirà la informació?) o físiques (quina grandària tindran les prestatgeries dels documents?) abans de plantejar les qüestions generals (quin és el propòsit de la base de dades?) o lògiques (quines entitats formaran part de la base de dades?). El quadre sinòptic següent sintetitza aquestes idees:

Quadre 1. Direcció del disseny en el procés de creació d'un sistema d'informació

- Del que es coneix al que es desconeix.
- Dels aspectes lògics als aspectes físics.
- D'allò general a allò concret.

Pel que fa al procés de cració, cadascuna de les tres fases enunciades abans (anàlisi, disseny, implantació) es pot dividir en totes les subfases que calguin segons el projecte concret i la classe de sistema que s'està dissenyant.

En el cas d'una base de dades documental, les dues primeres fases es poden subdividir en dues subfases (a i b). Les fases d'implantació es poden subdividir en cinc subfases (a, b, c, d i e). Novament cal indicar que aquestes divisions tenen sempre una part arbitrària. Aquí es fa una proposta concreta, però altres formes de dividir el procés de creació poden ser vàlides. En concret, en aquesta metodologia es proposa la divisió de fases del quadre següent:

Quadre 2. Procés de creació d'una base de dades documental

1. Anàlisi

1a. Anàlisi de l'empresa o organització (sistema d'activitats humanes) i del seu entorn.

1b. Anàlisi dels objectes candidats a ser registrats (sistema d'entitats registrables).

2. Disseny

2a. Disseny del model conceptual.

2b. Determinació del tractament documental (descripció, anàlisi i indexació documental, etc.).

3. Implantació

3a. Selecció del suport informàtic (*programari* i *maquinari*) d'acord amb els requeriments expressats en el model conceptual de la base de dades produït en la fase 2a i d'acord amb els requeriments expressats en la fase 2b.

3b. Elaboració del pressupost i del calendari d'implantació.

3c. Instal·lació, proves de rendiment i reelaboració, si escau, dels punts previs d'aquest procés de creació.

3d. Elaboració del llibre d'estil de la base de dades.

3e. Càrrega de dades, formació d'usuaris i promoció del producte.

Encara que, expressat en fases enumerades seqüencialment, el procés de creació sembla estrictament lineal, en realitat també té molt d'iteratiu, perquè tot i que sempre es comença per la fase d'anàlisi i es continua amb la de disseny, arribats a la fase 2b, per exemple, és possible que el dissenyador vulgui considerar de nou alguns aspectes de 2a, o que necessiti aclarir millor algunes qüestions d'1b, etc.

En aquest sentit, s'ha de fer notar que la metodologia no exclou totalment el procediment d'assaig i error, com ja s'ha advertit, sinó que l'integra d'una manera controlada per a refinar el producte.

En particular, és pràcticament impossible produir un model conceptual correcte en el primer intent, i l'experiència indica que el més probable és que el model elaborat en els punts 2a i 2b s'hagi de refer més d'una vegada, almenys en algun dels seus aspectes, principalment a la vista de les primeres proves de rendiment (3c).

Naturalment, ha d'arribar un moment en el qual el dissenyador doni per finalitzat el procés, però la qüestió de quantes vegades convé repetir-lo abans de donar-lo per bo no es pot establir *a priori,* sinó que més aviat és una qüestió sensible al context i que ha de decidir el dissenyador en cada cas.

En tot cas, és important que s'arribi a la fase d'implantació amb un model tan sòlid com sigui possible perquè a partir d'aquesta fase ja no resulta tan fàcil reconsiderar el projecte, si més no sense pagar algun preu, de manera que el

punt 3c s'hauria de considerar el punt d'enlairament, d'alguna manera, el punt de no tornada del projecte.

La fase d'implantació pot dur-la a terme un equip diferent del que va fer el disseny. De fet, en algunes empreses, sobretot en empreses mitjanes i grans, pot ocórrer que el departament d'informàtica s'encarregui de la fase d'implantació, encara que l'anàlisi i el disseny els hagi fet el de documentació. En empreses petites, el més habitual és que tot el procés l'executi un mateix equip o una mateixa persona.

Cadascuna de les fases precedents (anàlisi, disseny, implantació) té uns objectius, ha de produir uns resultats concrets i utilitzar unes eines determinades.

3.1. La fase d'anàlisi

L'objectiu d'aquesta fase és conèixer bé aquella part del món real, que denominem **sistema objecte**, que justifica i requereix la creació del sistema d'informació, d'una base de dades en aquest cas.

A l'efecte d'anàlisi, el sistema objecte es considera dividit en:

- **Un sistema d'activitats humanes (SAH):** l'empresa, l'organització, el sistema social, etc., que necessita o justifica la base de dades.

- **Un sistema d'entitats registrable (SER):** les coses, les persones o els conceptes que estaran representats en la base de dades.

Per tant, i atès que les característiques del sistema d'activitats humanes (SAH) determinaran les característiques de la base de dades, caldrà conèixer-les tan bé com sigui possible abans d'iniciar qualsevol activitat de disseny.

El sistema d'activitats humanes (SAH) es refereix a l'organització, l'empresa o, en termes generals, el sistema social –és a dir, un sistema format per persones i coses– que justifica o exigeix l'existència de la futura base de dades. En aquesta organització social desenvolupen les seves activitats els futurs usuaris que necessitaran que hi hagi un sistema d'informació (de vegades, ens pot convenir considerar que, al seu torn, dins del SAH, podem distingir entre el posseïdor o propietari del sistema i els usuaris o beneficiaris del sistema (Checkland, 1981)).

Per exemple, si pensem en l'OPAC (catàleg en línia) d'una biblioteca universitària com en un sistema d'informació, llavors el sistema objecte que modela és la universitat de la qual forma part, que necessita la biblioteca (així com altres recursos documentals) per a les seves activitats de creació i difusió del coneixement. En quin sentit, llavors, l'OPAC de la biblioteca modela d'alguna

manera la universitat? En el sentit en què el llenguatge documental amb el qual descriu els documents, la mateixa selecció dels documents que adquireix, els procediments de treball, els serveis que presta, etc., són un reflex de les característiques de la universitat.

Si considerem ara la base de dades d'una empresa periodística, la mateixa empresa periodística és el SAH del sistema objecte, però el públic interessat en la consulta d'aquesta base de dades formarà part també del SAH, en aquest cas, com a beneficiaris del sistema.

Atès que l'entorn sempre influeix en el sistema, de vegades de manera decisiva, els dissenyadors de la base de dades també hauran de conèixer les característiques de l'entorn de l'empresa (o del SAH, en termes més abstractes).

Al seu torn, el conjunt de coses, entitats o documents que haurà de ser descrit i representat en la base de dades forma l'anomenat *sistema d'entitats registrables (SER)*. Quan pensem en una base de dades documental és normal pensar en documents (p. ex., en documents impresos), però des d'un punt de vista abstracte això és inexacte. En primer lloc, en rigor, una base de dades conté representacions d'entitats i no necessàriament les entitats en si (penseu en una base de dades de patrimoni arquitectònic). En segon lloc, en una base de dades documental podem tenir els següents tipus d'*entitats representades*:

a) Coses: com documents en paper (bases de dades bibliogràfiques), pel·lícules (bases de dades de cinematografia), obres d'art (bases de dades de museus) o monuments (bases de dades de patrimoni arquitectònic).

b) Persones: dades biogràfiques de personatges històrics o de personalitats contemporànies, càrrecs de l'Administració, etc.

c) Conceptes: com idees i teories (bases de dades d'enciclopèdies i diccionaris).

Per tant, lluny de limitar-se a documents impresos com el seu únic objecte, les bases de dades documentals poden contenir representacions d'un nombre il·limitat de classes de coses. Aquestes possibles classes de coses susceptibles d'estar representades en una base de dades les denominem en l'argot tècnic *entitats*. Per tant, a més de considerar l'empresa (el SAH), en el procés de disseny hem de considerar també el conjunt d'entitats que haurem de representar en la futura base de dades (SER).

Els conceptes *objecte* i *entitat* es defineixen en l'apartat "Alguns conceptes bàsics" del mòdul "Sistemes de base de dades".

En el cas de la base de dades d'una empresa periodística, per continuar amb un exemple que ja hem esmentat, el SER consistirà en les informacions d'actualitat que publica aquesta empresa, sense perjudici d'altres tipus d'entitat. Per exemple, una de les agències de notícies més importants del nostre país, l'Agència EFE, produeix bases de dades no només sobre notícies d'actualitat sinó sobre personatges (biografies), sobre organismes i legislació de la Unió Europea (directori, disposicions legals, etc.).

Amb els dos principis fonamentals anteriors ja es disposa d'un aparell conceptual mínim que permet iniciar la discussió dels altres elements de la metodologia. S'observarà que algunes eines de l'aparell instrumental, com el *model entitat-relació* (que s'explica més endavant), també inclouen aspectes conceptuals. En realitat, en bona part és arbitrari decidir quins elements pertanyen a l'aparell conceptual i quins elements pertanyen al procedural o a l'instrumental. Aquí s'ha fet una elecció concreta, però probablement altres interpretacions són possibles.

El resultat d'aquesta fase d'anàlisi és una descripció textual que pot incloure, si cal, sobre el SAH que se sol denominar *informe de funcions* o *informe d'oportunitat,* i que ha d'incloure, com a mínim, els aspectes següents:

1) Propòsit i objectius de l'empresa o organització (SAH).
2) Propòsits i objectius de la futura base de dades o del sistema d'informació.
3) Identificació i característiques principals de les entitats registrables (SER).
4) Sistemes similars ja en funcionament, si escau.

L'eina principal aquí és la realització d'entrevistes amb representants de l'empresa o organització (SAH), així com l'anàlisi de qualsevol documentació sobre l'empresa que pugui aportar una comprensió global del sistema. Entre aquests documents podem citar organigrames, documents fundacionals, memòries, etc. Per descomptat, les entrevistes amb els futurs usuaris del sistema, així com amb la persona o els representants de l'empresa que hagin fet l'encàrrec, seran bàsiques.

En molts casos, ens podrem beneficiar d'un estudi de tipus *benchmarking* o de referenciació. Si ja hi ha altres bases de dades similars, serà convenient procedir a algun tipus d'estudi o anàlisi.

El resultat d'aquesta fase ha de consistir en la identificació clara i sense ambigüitats no només de les coses, les persones o els conceptes (entitats) sobre els quals la base de dades haurà de mantenir informació, sinó també de les funcions i els beneficis que s'esperen de la futura base de dades.

Exemple

Si l'encàrrec consisteix en el disseny d'una base de dades documental per a un museu, seria convenient programar visites a algun museu que ja en tingui. Si no hi ha altre remei, gairebé sempre podrem trobar models de bases de dades en funcionament a través d'Internet.

L'*informe de funcions* hauria de ser aprovat per la persona que fa l'encàrrec de la base de dades, com a manera d'assegurar-nos que totes dues parts, l'empresa i els dissenyadors de la base de dades, comparteixen les mateixes idees bàsiques.

3.2. La fase de disseny

El propòsit de la fase de disseny és obtenir un *model conceptual* de la base de dades i que també contingui una *proposta de tractament documental*. El primer element conté les indicacions necessàries per a orientar el procés d'implantació. El segon element estableix criteris i orientacions sobre el procés de des-

cripció i de representació del contingut semàntic de les entitats dels quals tractarà la base de dades.

Els dos components esmentats són el resultat de la fase de disseny i han de ser aprovats també per qui va encarregar el projecte, abans que puguin servir com a guies d'implantació. Per tant, el model conceptual no només ha de ser encertat, sinó que, a més, ho ha de semblar.

El *model conceptual* ha de contenir, almenys, els elements següents:

1) Objectiu i propòsits de la base de dades amb identificació dels usuaris del sistema (pot repetir parts essencials de l'informe de funcions, si és necessari).

2) Una definició dels àmbits o continguts temàtics de la base de dades.

3) Una identificació de les entitats representades en la base de dades.

4) El diccionari de dades.

5) Una descripció funcional que ha d'incloure els elements següents:

a) Quina classe d'informació es tractarà i com entrarà la informació en el sistema.

b) Quins processos documentals es duran a terme.

c) Quins serveis i productes generarà el sistema o a quines aplicacions podrà donar suport.

6) Una proposta de tractament documental.

L'àmbit *temàtic* de la base de dades és el conjunt dels temes o entitats sobre els quals la base de dades manté informació. Com tot àmbit, es pot definir per extensió o per comprensió. Per tant, pot ser tan breu com el nom d'una o més disciplines científiques; per exemple, l'àmbit de la base de dades LISA Plus són les *ciències de la documentació*. O pot consistir en una frase; per exemple, l'àmbit o contingut de la base de dades TESEO s'enuncia dient que està format per les *tesis doctorals publicades per universitats espanyoles*.

Atès el seu contingut, les eines per produir el document anterior són, entre altres, les següents:

1) l'informe de funcions elaborat prèviament,
2) el model entitat-relació i
3) el diccionari de dades.

3.2.1. El diccionari de dades

El diccionari de dades (*data dictionary*) és una eina que ajuda el dissenyador d'una base de dades a garantir la qualitat, la fiabilitat, la consistència i la co-

herència de la informació introduïda en la base de dades, i que condiciona decisivament, per tant, el rendiment i la qualitat global del sistema d'informació.

Consisteix en la llista detallada de cadascun dels camps de la base de dades amb l'especificació, per a cadascun d'ells, d'un conjunt de paràmetres que inclou, com a mínim, els aspectes següents:

1) Etiqueta.
2) Domini.
3) Tipus.
4) Indexació.
5) Tractament documental.
6) Llengua.
7) Altres controls de validació o observacions.
8) Obligatorietat.
9) Repetibilitat.
10) Instruccions per a l'entrada de dades.

Exemple

Suposem, a l'efecte d'aquesta explicació, una base de dades documental imaginària sobre notícies d'actualitat amb només tres camps: <Títol>, <Descriptors> i <Data de publicació>. El diccionari de dades tindria llavors aquesta forma (el diccionari de dades real tindria més camps):

Etiqueta: Títol

Domini:
Títol del document.
Tipus:
Alfanumèric.
Indexació:
Indexat.
Tractament documental:
Llenguatge lliure.
Llengua:
Llengua del document.
Controls de validació:
No pot quedar buit. Si, per alguna raó, el document manqués de títol, el documentalista assignarà un títol descriptiu.
Obligatorietat:
Obligatori.
Repetibilitat:
No és un camp repetible.
Instruccions per a l'entrada de dades:
Les diverses parts del títol es transcriuran de la manera següent: *títol: avanttítol: subtítol*. Els articles inicials no es posposaran. Per exemple, "Desacord a Brussel·les: reunió dels ministres d'economia: Es qüestiona el pacte d'estabilitat".

Etiqueta: Descriptors

Domini:
Els descriptors s'hauran d'obtenir del tesaurus de la base de dades.
Tipus:
Alfanumèric.
Indexació:
Indexat.
Tractament documental:
Llenguatge controlat.
Llengua:
Del centre de documentació.

Controls de validació:
No pot quedar buit i només admet valors extrets d'una llista de termes autoritzats.
Obligatorietat:
No.
Repetibilitat:
Sí. Es poden assignar diversos valors a aquest camp.
Instruccions per a l'entrada de dades:
S'assignaran descriptors (és a dir, termes d'indexació) que expressen els conceptes principals continguts en el document, segons el principi general següent: si l'article conté *n* conceptes rellevants s'assignaran *n* descriptors (fins a un màxim de 20 descriptors per document). Se seguiran les normes ISO/UNE de determinació de temes de documents i d'assignació de descriptors. Els termes se separaran amb coma, ",". Per exemple, edició òptica, publicació digital, documentació.

Etiqueta: Publicació

Domini:
La data de publicació de la notícia.
Tipus:
Data.
Indexació:
Indexat.
Tractament documental:
No escau.
Llengua:
No escau.
Controls de validació:
No admet valors fora d'interval.
Obligatorietat:
Sí.
Repetibilitat:
No.
Instruccions per a l'entrada de dades:
Les dades s'han d'introduir en el format: dd/mm/aaaa. P. ex. 28/11/2003.

Estudiant l'exemple de diccionari de dades anterior, format únicament per tres camps a l'efecte de la didàctica, podem observar quatre aspectes importants per al disseny de bases de dades:

1) Que el **domini**, en el context del diccionari de dades, es refereix al conjunt del qual un camp pot obtenir els seus valors.

2) Que el **tipus** es refereix, en canvi, al tipus de dada que admet el camp. Els tipus de dades solen ser: numèric, alfanumèric, data i lògic.

Recordem que un tipus de dada (*data type*) defineix un conjunt d'operacions vàlides i un interval de valors acceptable. Per exemple, el tipus de dades "alfanumèric" defineix operacions de comparació de cadenes de caràcters, entre altres, així com qualsevol lletra de la *a* a la *z* i qualsevol nombre del 0 al 9, així com qualsevol combinació d'aquests caràcters en paraules, frases, paràgrafs, etc. En canvi, no admet operacions aritmètiques, encara que admeti nombres. Per contra, un tipus de dada "numèrica" admet només nombres així com qualsevol operació aritmètica, etc.

Per la seva banda, un camp de dates només admet data en un format establert i permet cerques per intervals de data o per valors superiors o inferiors a una

data donada. Un camp lògic només admet un de dos valors: sí o no, veritable o fals.

3) Que el **tractament documental** estableix si s'ha d'utilitzar algun llenguatge documental per entrar els valors del camp, com succeeix en el camp Descriptors, en què el diccionari de dades estableix que aquest camp només admet paraules clau autoritzades extretes d'un tesaurus d'una llista d'autoritats.

Llista d'autoritats

Permet assegurar per a cada camp que qualsevol punt d'accés sigui únic i no es pugui confondre amb cap altre punt d'accés.

4) Que la **llengua** pot ser o bé la llengua del document o bé la del centre de documentació. Això significa, en el cas d'un document escrit en anglès, que el títol estaria en anglès però els descriptors estarien en castellà, sempre d'acord amb el diccionari de dades precedent.

A tall de síntesi, la taula següent recull els grups de camps que normalment trobarem en una base de dades de tipus documental. Quan fem el disseny del diccionari de dades, és aconsellable cotejar-lo amb aquesta taula i comprovar que no oblidem alguna categoria o algun grup de camps:

Taula 7. Grups de camps en una base de dades documental

Camps	Explicació
De control	Controlen la gestió interna del registre. Per exemple, el número del registre (ID), la data d'entrada, la data de modificació, etc.
Descriptius	Descriuen les característiques de les entitats o els documents de la base de dades (autor, títol, data, etc.).
Temàtics	Representen el contingut o tema del document o l'entitat representada en la base de dades (resum, descriptors, etc.).
Drets	Indiquen, si escau, quines restriccions o quins drets limiten la utilització del document o qui els posseeix.
Ubicació	Indiquen, si escau, la ubicació o localització del document original. Es poden referir a la ubicació física d'un document, (un llibre en una prestatgeria), o poden consistir en un punter informàtic que obre el document original en el cas de documents digitals.

3.2.2. ISBD i models canònics

ISBD

La ISBD (descripció bibliogràfica normalitzada internacional o *international standard bibliographic description*) és un estàndard que determina la forma i el contingut de la descripció bibliogràfica.

No hauríem d'oblidar que, en documentació, l'experiència prèvia ha deixat ben establerts quins són els atributs d'algunes entitats i, fins i tot, quina és la manera més convenient de representar-los. Podem parlar llavors de situacions canòniques que han generat un model. La millor eina d'anàlisi i de disseny, en aquest cas, consisteix precisament a aplicar aquest model ben conegut i provat.

Exemple

Els atributs estructurals de qualsevol classe de document poden ser adequadament modelats seguint les normes internacionals ISBD. Aquestes normes internacionals representen un gran esforç d'abstracció per proporcionar un marc general de descripció, vàlid per a qualsevol classe de document, des d'una partitura musical fins a una filmació audiovisual, passant per un arxiu d'ordinador, un fonograma o un article de revista, de manera que les ISBD constitueixen una eina de disseny de primera magnitud per a qualsevol problema documental on calgui de representar documents.

Sobre l'ús de les ISBD, cal advertir que alguns centres de documentació s'han sentit intimidats davant la complexitat aparent de la norma i la suposada obligació d'adoptar-la com un tot, inclosa la puntuació prolixa que prescriu i, en aquest sentit, s'ha argumentat que utilitzar la norma ISBD només té sentit en el context de les biblioteques normalitzades, aquelles que necessiten intercanviar registres i que, per tant, segueixen estàndards internacionals.

Entenem que aquesta postura és un error: primer, perquè sempre podem utilitzar l'estructura de les ISBD com una orientació en l'anàlisi dels documents convencionals així com una font d'inspiració per a situacions més exòtiques, independentment del fet que incorporem o no la norma en tota la seva complexitat, és a dir, incloent tots els nivells de descripció i totes les prescripcions de puntuació, principalment quan el fet de separar zones mitjançant camps allibera de la necessitat d'utilitzar la puntuació prescrita.

A més, en cas necessari, l'SGD hauria de permetre (com és el cas de diversos d'aquests programes; per exemple, Inmagic o CDS/ISIS) presentar la sortida de les dades en format ISBD (o en qualsevol altre format), des del moment en què l'estructura repetitiva dels registres permet incorporar instruccions del tipus: "el valor del camp *Títol* es transcriu seguit per un punt, un espai i un guió", etc.

3.3. La fase d'implantació

Una vegada aprovat el model conceptual de la base de dades, es pot procedir a la seva implantació, la qual pot seguir el procés següent:

1) Preselecció del sistema informàtic (programari + maquinari)

Tret que l'equip informàtic formi part de les restriccions inicials, probablement serà necessari examinar diversos programes candidats fins que hi hagi una certesa raonable que el programa triat s'ajusta bé als requeriments del model conceptual.

Per seleccionar el programa més adequat serà necessari posar-se en contacte amb diverses empreses del sector per sol·licitar documentació sobre les prestacions dels seus programes, els pressupostos, etc. Entre els criteris que ens ajudaran a prendre una decisió haurem de considerar els següents (a més d'altres criteris *ad hoc* segons la naturalesa específica del projecte):

a) Grau de compatibilitat amb la plataforma informàtica de l'empresa o corporació.

b) Grau de satisfacció dels requeriments establerts en el disseny conceptual.

c) Possibilitats de parametrització i disponibilitat d'eines de desenvolupament disponibles amb l'aplicació (llenguatge de guions, eines de programació addicionals, etc.)

Llenguatge de guions

Llenguatge de programació simple usat per l'usuari final que utilitza conjunts d'instruccions per lots (scripts) per a interactuar amb el 5.0, controlar diverses aplicacions, etc.

d) Valoració d'altres clients i usuaris. Base instal·lada de clients: poden proporcionar referències d'altres clients? Existeix un club d'usuaris de l'aplicació?

e) Utilització d'estàndards ben establerts, ja siguin *de facto* o *de iure,* i compatibilitat amb sistemes oberts (per exemple, compatibilitat amb el format PDF i amb el llenguatge HTML; o, en un altre extrem, compatibilitat amb l'ús de metadades i normes com Dublin Core, etc.).

f) Cost econòmic (pressupost).

L'ordre indicat no és significatiu: en alguns casos, el punt f) pot ser primordial mentre que el punt a) pot mancar d'importància, etc. En cada projecte concret, els responsables decidiran quin és l'ordre més adequat segons el context. En tot cas, els punts de a) a f) constitueixen un bon conjunt d'elements de partida que s'hauran de considerar en gairebé qualsevol projecte.

2) Elaboració del pressupost i del calendari d'implantació

a) Una vegada seleccionada una aplicació candidata, es procedeix a la instal·lació del programa i a una primera implantació de la base de dades aplicant el model conceptual creat en la fase de disseny per a fer les primeres proves.

b) Si s'aprova finalment l'ús de l'aplicació triada, es procedeix a la designació d'un administrador de la base de dades que, a partir d'ara, serà el màxim responsable d'aquesta i començarà a desenvolupar la primera versió de la base de dades segons s'indica en el punt següent.

3) Implementació dels controls terminològics; almenys d'aquella classe de controls dels quals es tingui la possibilitat d'instal·lació anticipada a la càrrega de dades (paraules buides, sinònims, valors predefinits, etc.).

4) Realització de proves amb una col·lecció de prova (conjunt de documents o d'entitats, candidats/tes a ser representats/ades) per a comprovar la consistència dels models i esquemes de registres detallats en les fases prèvies i continguts en el diccionari de dades.

5) Introducció de canvis o ajustos segons el resultat de les proves anteriors, si escau, en les definicions dels camps o en l'estructura del registre.

6) Automatització de processos repetitius: facilitats per donar altes (càrrega de dades), fer exportacions, fer consultes més freqüents, etc.

7) Segona càrrega de dades amb una altra col·lecció petita de documents, (per exemple, amb 15 o 20 documents). En aquest punt, és convenient simular tots els processos que es duran a terme amb aquesta base de dades: consultes, exportacions, etc. per obtenir la seguretat que es va pel bon camí. És normal que en aquesta fase apareguin imprevistos: formats d'exportació en els quals no s'havia pensat, tipus de consultes que requereixen algun reajustament en els camps, etc.

8) Prova d'usabilitat. Es pot aplicar ara una prova d'usabilitat i encarregar a diversos futurs usuaris reals de la base de dades (entre tres i cinc són un bon nombre) que facin proves d'ús realistes de la base de dades, encarregar-los tasques seleccionades prèviament per a aquest estudi i observar com les resolen. A més, també els demanarem que donin la seva opinió sobre el seu rendiment: és el que esperaven? Falta alguna opció? És fàcil d'usar?

S'incorporaran els canvis en el disseny si es detecta alguna insuficiència i, possiblement, ja haurem arribat al disseny definitiu (en tot cas, no podem estar modificant el disseny de manera indefinida i tard o d'hora haurem de donar per bo el model).

9) Disseny de les vistes dels usuaris i de les caràtules de portada o inici.

10) Definició dels grups d'usuaris i d'altres responsables de la base de dades. Cada grup d'usuaris ha de tenir privilegis i, si pot ser, vistes diferents de la base de dades.

En aquest sentit, és convenient considerar que hi sol haver, almenys, quatre tipus de persones involucrats en la base de dades, que són els següents:

a) L'**administrador** o director de la base de dades. És la persona que té la màxima responsabilitat en la base de dades. Aquesta figura ja ha estat designada abans.

b) Els documentalistes/analistes. Són els que duen a terme l'anàlisi de la informació. Solen ser professionals experts en la temàtica de la base de dades que produeixen resums o assignen descriptors.

c) Els operadors. Són els que fan la càrrega de dades. Operadors i analistes poden ser les mateixes o diferents persones. En aquest últim cas, de vegades, els analistes fan la seva feina sobre plantilles que després els operadors bolquen en la base de dades.

d) Els usuaris finals. Són els que explotaran i utilitzaran la informació. Hi pot haver diverses categories d'usuaris finals. Per exemple, a Internet és habitual que hi hagi usuaris que només poden fer consultes, però no poden veure els resultats complets, tret que estiguin registrats o que abonin una quota, etc.

11) Inici del procés de càrrega de dades sitemática i d'explotació del sistema.

Com és lògic, arribarà el moment en què haurem de començar la càrrega de dades de manera massiva i sistemàtica. Per a això, haurem d'establir de manera explícita, clara i sense ambigüitats els extrems següents:

a) Rutines de càrrega de dades. Qui, com i quan es fa la càrrega de dades. Els encarregats de fer-la han de ser personal format no solament en l'ús del programa, sinó també en el coneixement del diccionari de dades.

b) Rutines de seguretat. Qui, com i quan es creen les còpies de seguretat. En tot cas, s'haurien de fer almenys dues còpies de seguretat i en dos formats diferents. Una còpia de seguretat dels treballs del dia i una altra, desfasada respecte a l'anterior en un o més dies.

És convenient tenir còpies de seguretat en dos formats: el format natiu del programa més un format fàcilment explotable amb altres aplicacions o bases de dades. El més fàcil és tenir una còpia de seguretat en format ASCII i en un format tipus "camps separats per tabulador", que entenen molts programes de base de dades.

c) Avaluació i controls de qualitat. Periòdicament establirem controls de qualitat. Els elements més típics en aquest control són: el control de duplicats i el control de la qualitat de la indexació. Per a tots dos tipus de controls, les bases de dades documentals solen proveir opcions diverses. En molts programes documentals, per exemple, podem exportar i publicar la llista de descriptors i revisar-la periòdicament. També podem sol·licitar la detecció de duplicats. L'apartat següent es dedica a analitzar, de manera detallada, tot el que es refereix a avaluació i control de qualitat en bases de dades.

A més, segons la naturalesa de la base de dades, establirem altres tipus de controls adequats al seu contingut, etc.

d) Política de manteniment i explotació. S'editarà la versió 1 del llibre d'estil de la base de dades, que inclou:

- Versió definitiva del model conceptual.
- Normativa de tractament documental.
- Política de formació del personal tècnic i organització de sessions de formació dels usuaris finals.

12) Accions de promoció, si escau.

3.4. Conclusions

El valor d'aquesta metodologia rau, com ja s'ha dit al principi, en el fet que ajuda a fer que el producte final sigui més resultat del disseny conscient que de

les forces cegues de l'atzar o de l'assaig i error, però, particularment entenem que la seva utilitat augmenta quan s'aplica a situacions poc canòniques o a situacions atípiques, com les que l'entorn canviant de la nostra professió introdueix en cada moment i, pel que sembla, tal com el nou horitzó de les autopistes de la informació i d'un futur món digital sembla prometre.

Esperem que, llavors, l'aplicació d'aquesta classe de metodologies serveixi perquè els professionals d'aquest camp puguin demostrar els beneficis d'una formació acadèmica adequada, del treball ben fet i de la planificació, perquè en aquest camp d'activitats també és rigorosament cert que l'èxit es deu invariablement a "un deu per cent d'inspiració i un noranta per cent de transpiració".

4. Avaluació de bases de dades

En els apartats precedents s'han tractat qüestions diverses referides al disseny, la producció i la distribució de bases de dades documentals que haurien de servir per a fer un producte concret, és a dir, coses com ERIC, la base de dades d'articles de revistes d'educació, l'arxiu de premsa d'*El País*, la base de dades de fotografies AGE Fotostock o la base de dades de recursos d'Internet Intute.

Ara bé, una vegada aquestes bases de dades es troben en el mercat, les qüestions són: és una base de dades competitiva? En quina posició es troba en el mercat respecte d'altres productes similars? Satisfà els requisits mínims de qualitat? Aquest tipus de qüestions, que interessen tant a l'usuari o client de la base de dades com al seu productor, constitueixen l'eix d'aquest apartat.

La nostra aportació consisteix a presentar una aproximació que procura ser global i integradora i que es compon de dues parts. En la primera, més extensa, es recopilen i s'organitzen els indicadors o criteris principals que han estat considerats en els estudis fets fins al present per a l'avaluació de la qualitat de bases de dades i que, per tant, no se circumscriu tan sols als sistemes de recuperació (SR) sinó que també inclou el contingut, és a dir, la mateixa base de dades. En la segona, es descriuen les principals tècniques de recollida de dades sobre l'usuari, un element essencial per poder mesurar aquells criteris d'avaluació que es refereixen especialment a aspectes subjectius de l'usuari (satisfacció, utilitat, etc.).

D'altra banda, cal recordar que aquest tipus d'estudis d'avaluació han pres com a objecte bases de dades de tipus molt diferents. En primer lloc, els catàlegs de biblioteca en línia, per als quals es disposa d'una normativa de caràcter internacional que persegueix facilitar l'intercanvi de registres; a continuació les bases de dades cientificotècniques, amb caràcter especialitzat, per a les quals, en canvi, no es disposa de normes de seguiment comú i cada productor aborda com ha cregut més convenient; i finalment els serveis de cerca a Internet. Un repàs dels estudis d'avaluació en bases de dades ens mostra com en el curs dels anys s'han anat presentant anàlisis diferents d'aquests objectes. L'enfocament que presentem vol tenir un caràcter integrador i, per tant, és aplicable tant a catàlegs de biblioteques com a bases de dades especialitzades o motors de cerca.

4.1. Indicadors (criteris d'avaluació)

Per establir una metodologia d'avaluació i anàlisi de la qualitat de les bases de dades, cal determinar, primer de tot, quins seran els indicadors que es tindran en compte. Aquest és un requisit imprescindible per a qualsevol tipus d'avaluació

que es vulgui dur a terme sobre qualsevol servei o producte (per exemple, avaluació de revistes, avaluació de sistemes de recuperació de la informació, etc.).

Els antecedents són diversos. Podem començar fent referència al treball realitzat pel Southern California Online User Group (SCOUG) que, el 1990, en el marc del *Fourth Annual Retreat* dedicat a la medició de la qualitat de bases de dades (*measuring the quality of databases*) va establir deu criteris per a l'avaluació que han estat la base de molts estudis posteriors, com els de Wilson (1998), Xie (1998) o Rodríguez Yunta (1998), entre molts altres, que inclouen propostes de les quals es pot extreure un nombre important d'indicadors o criteris d'avaluació. La popularització dels cercadors web a la fi dels anys 90 i principis del segle XXI matisa la importància d'alguns d'aquests criteris, que estan més aviat pensats per al context de les bases de dades cientificotècniques, encara que en línies generals són fàcilment adaptables a aquest context.

A fi d'organitzar una mica la llarga llista d'indicadors a la qual s'ha de fer referència, els hem agrupat en tres grans àmbits:

1) La **base de dades (el contingut):** inclou tot el que es refereix a la qualitat de la informació continguda en la base de dades, la seva indexació, els documents escollits, etc.

2) El **sistema de recuperació (el continent):** inclou tot el que es refereix al programa de recuperació i les seves prestacions i també a la interfície de consulta (no es pot oblidar que una mateixa base de dades pot estar disponible en diferents sistemes de recuperació de la informació).

3) La **gestió i administració de la base de dades:** inclou tot el que es refereix a la documentació sobre la base de dades, els procediments de cobrament, els preus i les facilitats atorgades als usuaris.

Taula 8. Criteris d'avaluació

Contingut de la base de dades	Grau d'exactitud i precisió Errors gramaticals i mecanogràfics Errors d'omissió Fiabilitat de les dades Registres duplicats Abast i cobertura Grau de cobertura o abast temàtic Cobertura geogràfica i lingüística Grau d'inclusió Estructura Grandària Nivell de creixement Actualització Grau d'actualització Període d'actualització Consistència Consistència de la catalogació Consistència en l'anàlisi de contingut

Sistema de recuperació	Prestacions del llenguatge d'interrogació Precisió Exhaustivitat Temps de resposta Utilitat Formats de visualització Usabilitat de la interfície
Gestió de la base de dades	Documentació sobre la base de dades Atenció a l'usuari Preu i sistema de facturació Sistema de distribució

4.1.1. Contingut de la base de dades

En aquest apartat s'avalua la matèria primera fonamental per assegurar la qualitat d'una base de dades. De poc servirà disposar d'un SRI amb moltes prestacions o d'una gestió i administració molt eficaç si els continguts són pobres i de poca qualitat. Jacsó (1997) presenta un article valuós de revisió bibliogràfica que repassa les publicacions principals que es refereixen als aspectes de qualitat del contingut: precisió i fiabilitat, abast i cobertura, actualització, etc.

L'avaluació del contingut comença quan el productor de la base de dades selecciona i analitza la informació i afecta aspectes que són de la seva competència directa.

1) Grau d'exactitud i precisió

Es tracta d'un conjunt d'indicadors que fan referència a problemes relacionats amb la falta de precisió en l'entrada de dades: errors gramaticals i mecanogràfics de les paraules, absència d'informació en els camps de la base de dades, fiabilitat de les dades o duplicació d'informació. En definitiva, agrupa un conjunt de qüestions relacionades amb la "qualitat de les dades", que podria ser una altra manera de dir el mateix.

a) Errors gramaticals i mecanogràfics

Es refereix a la presència d'errors ortogràfics, sintàctics (problemes de concordança de gènere, nombre, cas, etc. de les paraules) i de mecanografia.

Els errors gramaticals i de mecanografia afecten directament la recuperació de la informació, ja que podem deixar d'obtenir una informació pertinent o recuperar-ne una altra que no ho sigui. Són especialment preocupants en àrees com les finances o la informació mèdica o jurídica, en les quals es poden prendre decisions molt importants sobre la base del contingut de la informació recuperada.

b) Errors d'omissió

Els registres incomplets (p. ex. falta la data de publicació, l'idioma, el tipus de document, etc.) constitueixen errors d'omissió i es poden detectar i prevenir fàcilment. Per detectar-los, es poden usar en rutines automàtiques.

c) Fiabilitat de les dades

Es refereix a la correspondència exacta del contingut dels registres amb els documents als quals representen. En les bases de dades de recursos web, cal estar atents a la presència d'enllaços inexistents o sense actualitzar.

d) Registres duplicats

L'origen d'aquest problema són els errors d'exactitud que s'han descrit en aquest mateix apartat i també les inconsistències (vegeu el punt 4: "Consistència"). A l'efecte de la recuperació, l'existència de registres repetits dificulta la consulta de la base de dades perquè augmenta innecessàriament el nombre de resultats.

2) Abast i cobertura

a) Grau de cobertura o abast temàtic

Tota base de dades està especialitzada en una o diverses àrees temàtiques. La cobertura indica la proporció de fonts d'aquesta matèria concreta que estan disponibles en la base de dades. El sistema que es pot utilitzar per a mesurar aquest indicador consisteix a determinar la proporció de fonts d'informació considerades de màxim interès per a una àrea temàtica que forma part del conjunt de fonts d'informació buidades per la base de dades.

b) Cobertura geogràfica i lingüística

Tenint en compte a l'àmbit geogràfic i les llengües, es pot valorar l'internacionalisme de la base de dades, poc destacable en les anglosaxones. Per a l'usuari no anglosaxó aquest factor acostuma a tenir un valor notable, ja que també li interessa localitzar documents en el seu idioma.

c) Grau d'inclusió

Es refereix a la presència de determinats tipus de document: només articles de revista o també monografies, congressos, patents, normes, etc.

d) Estructura

Es refereix al nombre de camps definits i recuperables. Cal veure si només afecten la part descriptiva (autor, títol, etc.) o també el contingut (descriptors, classificació, resum, etc.).

e) Grandària

El nombre de registres (o la quantitat de pàgines web) indexats és un criteri del qual acostumen a presumir els grans productors de bases de dades i que impressiona de manera notable els usuaris. Encara que es tracta d'un indicador important, cal interpretar-lo de manera adequada i amb la perspectiva correcta. De què ens serveix tenir milions de registres o de pàgines web si no s'inclouen les fonts més prestigioses?

f) Nivell de creixement

Es mesura el nombre de registres nous per any. Aquest nombre hauria de ser igual o semblant al nombre de documents produïts de les matèries que tracta la base de dades en el mateix període de temps.

3) Actualització

a) Grau d'actualització (o actualitat de la informació)

Es mesura el temps que passa entre el moment en què un document està disponible i la seva inclusió en la base de dades.

b) Període d'actualització

Es refereix a la periodicitat amb la qual s'actualitzen els registres de la base de dades.

4) Consistència

Es refereix a la característica o propietat que posseeixen els registres d'una base de dades que estan confeccionats uniformement. Per aconseguir-ho és necessari aplicar estrictament i homogèniament un conjunt de normes comunes.

a) Consistència de la catalogació (o de la descripció)

Mesura el grau de coherència pel que fa a l'anàlisi formal dels documents, és a dir, pel que fa a la seva descripció bibliogràfica (assignació correcta de camps i subcamps) i l'elecció de punts d'accés, que impliquen mésproblemes en la recuperació. Els termes que constitueixen punts d'accés fonamentals als registres (autors, títol de la revista, matèria, etc.) han d'estar normalitzats, és a dir, s'han d'assignar de manera homogènia i consistent perquè, d'una altra manera, limiten les capacitats de recuperació.

La utilització de llistes de validació (domini del camp) és molt útil per assegurar la consistència en les entrades d'autor, títol de revista, matèria, etc. D'altra banda, és fàcil preparar proves de consistència de la informació que s'ha introduït en un

mateix camp. La més senzilla és generar els índexs del camp perquè, d'aquesta manera, es pugui comprovar si el control del domini s'ha fet correctament.

b) Consistència en l'anàlisi de contingut

Es refereix a la coherència en l'assignació de termes d'indexació i de codis de classificació per assegurar que s'utilitzen sempre els mateixos quan volem representar una temàtica idèntica.

4.1.2. Sistema de recuperació (o SR)

La major part dels elements que es poden portar a col·lació en aquest apartat ja han estat analitzats en els apartats 1 i 2.3. d'aquest mateix mòdul. Per això, en farem un repàs breu. Aquests indicadors estan relacionats amb el procés de cerca i visualització dels resultats i depenen, per tant, de les característiques del programa informàtic utilitzat.

De fet, aquests criteris són totalment independents dels anteriors. Fins i tot, es dóna el cas de bases de dades que són consultables per diferents SR, amb la qual cosa, si s'avaluen, cal precisar quin és el programa de recuperació que es té en consideració.

Bases de dades amb programes diferents

Medline en pot ser un exemple. El mateix contingut és accessible a la National Library of Medicine (www.nlm.nih.gov) i també al portal brasiler Scielo (www.bireme.br/bvs/e/ebd.htm), amb programes informàtics diferents.

1) Llenguatge d'interrogació

Les funcionalitats principals a considerar són l'ús d'operadors booleans, la cerca per camps, els operadors de proximitat, la possibilitat de mostrar índexs de camps, la possibilitat de mostrar i consultar el tesaurus, la cerca en llenguatge natural, cerques semàntiques, etc.

2) Precisió

Mesura la capacitat de l'SR per proporcionar tan sols els documents rellevants a la pregunta formulada per l'usuari. Els problemes o errors poden ser deguts tant a la imprecisió de la consulta com a la inconsistència en l'anàlisi.

3) Exhaustivitat

Es tracta de la proporció de documents rellevants que se subministren en resposta a una petició determinada respecte del total de documents que hi ha en la base de dades.

4) Temps de resposta

Mesura el lapse de temps transcorregut des de la formulació de la pregunta fins a l'obtenció de resultats. En alguns casos, no és fàcil de comptabilitzar, ja que la intensitat del trànsit a la xarxa és molt variable.

5) Utilitat

Es pot mesurar objectivament analitzant la consistència dels resultats, el grau d'actualització, la presència de duplicats, la proporció d'enllaços erronis o inexistents, etc., encara que no deixa de ser un indicador una mica subjectiu (doncs depén de la satisfacció de l'usuari respecte els resultats).

6) Formats de visualització

Es refereix a la possibilitat de seleccionar diferents formats ajustats a les necessitats dels usuaris. Inclou: Format breu per visualitzar la informació global, format ampli, amb resum, per poder seleccionar els registres concrets, etc. També es refereix a prestacions d'impressió, enregistrament o enviament per correu electrònic dels registres.

7) Usabilitat de la interfície

L'objectiu perseguit és la màxima usabilitat, és a dir, una presentació clara, senzilla, intuïtiva, etc. dels continguts de la base de dades i de les prestacions per a consultar-les.

Interfície de consulta

Els elements que formen part de la interfície d'interacció de l'usuari han estat resumits i descrits en el subapartat "Interfície de consulta" dins l'apartat "Distribució de bases de dades" d'aquest mateix mòdul: diversitat del sistema de consulta o de cerca adequada a usuaris experts i principiants (seqüencial, índexs, assistida, llenguatge d'interrogació), navegació, selecció d'idioma, sistemes d'ordenació de resultats, etc.

4.1.3. Gestió de la base de dades

Els criteris que s'indiquen a continuació mesuren l'eficàcia i el grau de qualitat del distribuïdor de la base de dades o del departament encarregat del màrqueting i la promoció.

Els distribuïdors tradicionals de bases de dades (com seria el cas de Dialog) constitueixen els exemples de més qualitat en aquest aspecte. Google també posa a l'abast de l'usuari una informació molt completa i detallada sobre l'estructura i les característiques del seu contingut.

1) Documentació sobre la base de dades

Es tracta d'avaluar si existeix una descripció clara i detallada de la base de dades i del sistema de consulta.

2) Atenció a l'usuari

Es té en compte l'existència de cursos de formació adreçats a diversos tipus d'usuari, o d'un servei més o menys permanent. Això tan sols es troba en el sector de les bases de dades cientificotècniques, ja que els grans serveis de cerca a Internet no es poden ni plantejar oferir un servei d'aquestes característiques als seus milions d'usuaris.

3) Preus i sistema de facturació

Aquest criteri es refereix a la relació qualitat-preu i els sistemes de facturació establerts per la base de dades. En molts casos, és difícil valorar i comparar, ja que els sistemes de cobrament utilitzats són complexos i no sempre tenen en compte els mateixos paràmetres (els registres visualitzats, o els baixats, el temps, etc.).

4) Sistemes de distribució

S'analitza si hi ha diversitat de sistemes: web i suport òptic són els canals principals. Es tracta d'un criteri amb poc pes, ja que el web s'ha convertit en el sistema de distribució per excel·lència.

4.2. Com avaluar la base de dades?

Els indicadors esmentats abans es poden mesurar mitjançant un sistema d'avaluació o de quantificació basat en anàlisis externes (funcionament del sistema, característiques del contingut de la base de dades, anàlisi dels registres, etc.). Ara bé, en la seva gran majoria també es poden avaluar des del punt de vista de l'usuari. Així doncs, sembla clar que el temps de resposta és un criteri que es pot mesurar objectivament comptant el lapse de temps transcorregut des que es demana executar una petició d'informació fins que apareix en pantalla una llista amb els resultats.

Per conèixer aquests valors es necessita utilitzar alguna tècnica específica de recollida de dades, ja sigui directa o indirecta. Les més conegudes són els qüestionaris i les entrevistes, dedicades a conèixer la satisfacció de l'usuari respecte de l'ús de la base de dades. També es pot fer ús de l'observació, normalment gravant les accions dels usuaris. En els últims anys està cobrant un interès notable l'anàlisi de transaccions (o de *logs*). Els qüestionaris i les entrevistes permeten conèixer de manera directa el que pensa l'usuari mentre que l'observació i l'anàlisi de transaccions permeten una aproximació indirecta, ja que tan sols permeten seguir les accions que l'usuari ha dut a terme però en desconeixen el context (la pregunta que es formula) i les seves impressions (satisfacció, utilitat, etc.).

4.2.1. Qüestionaris i entrevistes

Es tracta de dues tècniques que recullen les dades directament de l'usuari i que són molt conegudes i utilitzades en amplis àmbits d'investigació. En el tipus d'aplicacions a les quals fem referència tenen per objectiu determinar els coneixements, les opinions o les actituds dels usuaris respecte a les bases de dades que han consultat (per exemple, el seu grau de satisfacció).

La diferència entre totes dues tècniques és una mica subtil, ja que en tots dos casos es parteix d'un qüestionari previ més o menys estructurat; el que passa

és que en el qüestionari pròpiament dit les respostes són escrites per l'enquestat, i en l'entrevista, per l'enquestador, que formula les preguntes oralment. D'altra banda, el qüestionari permet recollir dades de grups més nombrosos de persones.

L'avantatge principal d'aquestes tècniques de recollida de dades rau en el fet que permeten un coneixement més profund de l'opinió i el grau de satisfacció de l'usuari que el que ofereixen mètodes indirectes com l'anàlisi de transaccions, ja que es demanen directament les seves opinions. Per contra, es tracta de sistemes lents (no es poden automatitzar) i cars (s'han de passar personalment) i no es poden administrar a un conjunt molt gran d'usuaris (especialment, l'entrevista).

4.2.2. Observació

L'observació, com a tècnica de recollida de dades, consisteix a prendre nota o registrar el desenvolupament d'una activitat durant un període de temps determinat. Es tracta d'efectuar una vigilància directa i un registre de les dimensions del fenomen que s'estudia (en el nostre cas, la consulta a una base de dades o, més genèricament, el comportament en el procés de cerca d'informació). En el context de la recuperació d'informació s'acostumen a utilitzar sistemes d'enregistrament (normalment, vídeo).

Es tracta d'una tècnica que aporta més objectivitat que l'entrevista o el qüestionari i que pot complementar la percepció subjectiva de l'usuari. D'altra banda, pràcticament no incomoda el subjecte observat i es pot aplicar en situacions en les quals els usuaris no són capaços de respondre adequadament un qüestionari (per exemple, nens, persones amb poca formació, etc.).

Ara bé, és una tècnica que requereix molta paciència, ja que la recollida de dades pot ser llarga i lenta, amb molts temps morts. També té un cert grau de superficialitat, perquè no proporciona una visió profunda (causes, etc.) del problema a tractar. Finalment, cal tenir en compte que la deontologia obliga a obtenir el permís de les persones estudiades.

4.2.3. Anàlisi de transaccions

L'anàlisi de transaccions (*transaction log analysis* o TLA) és una tècnica de recollida de dades que **registra** les accions dutes a terme per un usuari en un sistema de recuperació de la informació. La designació en anglès inclou la paraula *log* (registre), que evoca el registre cronològic de les operacions de processament de dades en un sistema que es registren en un fitxer.

L'anàlisi de transaccions s'ha utilitzat de manera extensiva des dels anys vuitanta per avaluar sistemes de gestió de biblioteques (la part pública, els OPAC) i és molt utilitzada.

Recuperació d'informació en el web

Jansen i Pooch (2001) fan una revisió bibliogràfica sobre els diversos treballs que s'han publicat sobre estudis de recuperació d'informació en el web i han mostrat com la gran majoria utilitza l'anàlisi transaccional com a base per a l'estudi.

L'estructura estàndard d'un **fitxer de registre** acostuma a incloure els elements següents: data i hora, identificador d'usuari, expressió de cerca (termes de consulta i operadors) i durada de la connexió.

Els analistes diferencien una sessió (entesa com el període de temps comprès des del moment en què l'usuari es connecta a una base de dades fins que l'abandona) d'una consulta (que és una part d'una sessió i que es refereix a l'expressió de cerca que l'usuari formula al sistema).

Expressió de cerca

Està formada per un terme o un conjunt de termes units, o no, per algun operador.

L'anàlisi de transaccions es pot fer sobre elements o indicadors diferents: durada (de la sessió i de la consulta), termes utilitzats, operadors booleans, camps utilitzats, nombre de documents recuperats, accions dutes a terme (impressió, exportació), etc.

Operadors booleans

Són operadors d'àrea basats en la lògica com AND, OR, NOT, etc.

Es tracta de la tècnica més simple per recollir dades sobre la interacció usuari-SR a distància, sense necessitat de presència humana externa i, per tant, sense molestar ni condicionar les accions de l'usuari. A més, es pot aplicar a un gran nombre d'usuaris, tal com es pot comprovar llegint alguns estudis que manegen milions d'interaccions.

Com en el cas de l'observació, és un estudi indirecte i una mica superficial que tan sols permet conèixer les accions dutes a terme per l'usuari, però sense saber res de les seves percepcions, les opinions (quina valoració fa del sistema o dels registres obtinguts), els coneixements previs o quina és la necessitat d'informació que vol satisfer. D'altra banda, els fitxers que es generen són molt voluminosos i és una mica difícil treballar-hi.

Com a valoració global podem dir que es tracta d'una tècnica que resulta limitada per a l'anàlisi del comportament dels usuaris. Malgrat això, les dades que proporciona poden ser molt útils per a fer propostes de millora de l'accés a la informació en un SR.

4.3. Conclusions

L'avaluació de bases de dades és un procés que interessa tant a usuaris com a productors. Als usuaris els ajuda a seleccionar els continguts més interessants i complets, utilitzar els millors SR o aprofitar els millors preus. Per als productors, l'interès per l'avaluació i la qualitat té un abast molt més profund, ja que una preocupació constant per aquestes qüestions els permetrà disposar d'un producte molt més competitiu. Si es disposa d'uns criteris o indicadors, es pot procedir a fer anàlisis periòdiques del grau de qualitat de la base de dades.

Aquestes anàlisis recolliran dades de caràcter objectiu sobre la base de dades i també s'interessaran per recollir, amb les tècniques directes o indirectes que han estat descrites, la utilitat i el grau de satisfacció dels usuaris.

A partir dels resultats de l'anàlisi, poden sorgir determinades propostes de canvi que podran afectar diferents aspectes de la gestió i el manteniment de la base de dades, ja siguin canvis al programa de recuperació de la informació, en l'adequació dels manuals i de les ajudes en línia o en el mètode de treball establert. Veiem, per tant, com les modificacions poden afectar qualsevol dels tres nivells analitzats: la base de dades (contingut), el sistema de recuperació (programa) o la gestió i administració.

En general, les empreses i els organismes productors de bases de dades haurien de considerar els elements discutits aquí com a part dels seus procediments de qualitat. Si les empreses que produeixen les bases de dades apliquen amb una periodicitat determinada procediments de control de la qualitat en altres àmbits, per què no fer-los extensibles a les bases de dades del seu departament de documentació?

Això encara és més necessari per a les empreses o els organismes l'activitat dels quals depèn en part (o totalment) de la qualitat de les seves bases de dades.

Bibliografia

Abad, F. (1997). *Investigación evaluativa en Documentación: aplicación a la documentación médica*. València: Universitat de València.

Abadal, E. (2002, setembre-octubre). "Elementos para la evaluación de interfaces de consulta de bases de datos" [article en línia]. *El profesional de la información* (vol. 11, núm. 5, pàg. 349-360). Swets & Zeitlinger Publishers. [Data de consulta: 25 de maig de 2010] (http://www.elprofesionaldelainformacion.com/contenidos/2002/septiembre/3.pdf)

Abadal, E.; Codina, L. (2005). *Bases de datos documentales: características, funciones y método*. Madrid: Síntesis.

Blair, D. C. (1990). *Language and representation in information retrieval*. Amsterdam [etc.]: Elsevier.

Checkland, P. B. (1981). *Systems thinking, systems practice*. Chichester: Wiley.

Codina, L. (1998). "Metodología de análisis de sistemas de información y diseño de bases de datos documentales: aspectos lógicos y funcionales". *Anuari SOCADI de documentació i informació* (pàg. 195-209). Barcelona: SOCADI.

Codina, Lluís (1993). *Sistemes d'informació documental: concepció, anàlisi i disseny de sistemes de gestió documental amb microordinadors*. Barcelona: Pòrtic.

Connolly, T. M. *et al.* (1995). *Database systems: a practical approach to design, implementation and management*. Wokingham: Addison-Wesley.

Hearst, M. A. (1999) "User interfaces and visualization". Baeza-Yates, Ricardo; Ribeiro-Neto, B. *Modern information retrieval* (pàg. 257-323). NovaYork: ACM; Harlow: Addison-Wesley.

Jacsó, P. (1997). "Content evaluation of databases". *ARIST* (vol. 32, pàg., 231-267).

Jansen, B. J.; Pooch, U. (2001)."A review of web searching studies and a framework for future research". *JASIS* (vol. 52, núm. 3, pàg. 235-246). Nova York: John Wiley.

Marchionini, G. (1995). *Information seeking in electronic environments*. Cambridge: Cambridge University.

Miguel, A. de; Paittini, M. G. (1999). *Fundamentos y modelos de bases de datos* (2a. ed.). Madrid: Ra-Ma.

Morville, P.; Rosenfeld, L. (2006). *Information architecture for the world wide web* (3a. ed.). Sebastopol (Califòrnia) (etc.): O'Reilly.

Nielsen, J. (2000). *Usabilidad: diseño de sitios web*. Madrid [etc.]: Prentice Hall.

Nielsen, J.; Loranger, H. (2006). *Usabilidad: prioridad en el diseño web*. Madrid: Anaya Multimedia.

Peña, R. (2002). *Gestión digital de la información: de bits a bibliotecas digitales y la web*. Madrid: Ra-Ma.

Information Market observatori (IMO) (1995). *The quality of electronic information products and services*. Imo Working paper 95/4.

Raya, F. (1987). *Database design for information retrieval: a conceptual approach*. Nova York [etc.]: John Wiley & Sons.

Rodríguez Yunta, L. (1998). "Evaluación e indicadores de calidad en bases de datos". *Revista española de documentación científica* (vol. 21, núm. 1, pàg. 9-23). Madrid: Consejo Superior de Investigaciones Científicas : Instituto de Información y Documentación en Ciencia y Tecnología.

Rodríguez Yunta, L.; Tejada, C. (coord.) (2003). *Directorio español de software para la gestión bibliotecaria, documental y de contenidos*. Madrid: Consejo Superior de Investigaciones Científicas.

Shneiderman, B.; Byrd, D.; Croft, W. B. (1997, gener). "Clarifying search: a user-interface framework for text searches" [article en línia]. *D-Lib Magazine* (vol. 3, núm. 1). [Data de

consulta: 25 de maig de 2010.]
<www.dlib.org/dlib/january97/retrieval/01shneiderman.html>

Soergel, D. (1985). *Organising information principles of data base and retrieval systems*. San Diego [etc.]: Academic Press.

Tramullas, J.; Olvera, M. D. (2001). *Recuperación de la información en internet*. Madrid: RaMa.

Van Rijsbergen, C. J. (1975). *Information retrieval* [en línia]. Londres: Butterworths. [Data de consulta: 1 de juny de 2010.]
<http://www.dcs.gla.ac.uk/Keith/Preface.html>

Villanueva, E. (1996, gener-juny). "Bases de datos y bibliotecología: cómo deshacer la innecesaria incomunicación". *Investigación bibliotecológica* (vol. 10, núm. 20, pàg. 27-32). Mèxic, DF: Centro Universitario de Investigaciones Bibliotecológicas, UNAM.

Willitts, J. (1992). *Database design and construction: an open learning course for students and information managers*. Londres: Library Association.

Wilson, T. D. (1998). "EQUIP: a european survey of quality criteria for the evaluation of databases". *Journal of information science* (vol. 24, núm. 5, pàg. 345-357). Amsterdam: Elsevier.

Xie, M.; W., H.; Goh, T. N. (1998). "Quality dimensions of Internet search engines". *Journal of information science* (vol. 24, núm. 5, pàg. 365-372). Amsterdam: Elsevier.