Bollicine di champagne...

Strumenti per la ricerca ed il recupero dell'informazione su Internet

di Gabriele Lunati

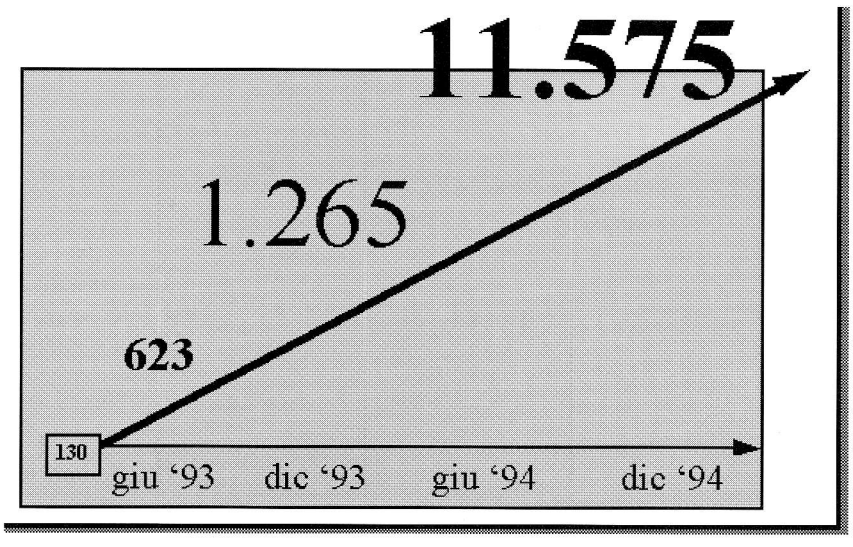
Web sites come and go like bubbles in champagne
John Markoff

Che il mondo fosse pieno di grafomani, soprattutto chi lavora in biblioteca l'aveva capito da un pezzo. La cosa oggi è sotto gli occhi di tutti vista la facilità con cui — grazie a Internet — qualunque istituzione o persona può creare, rendere disponibile e mantenere una pagina web: un fenomeno che è diventato esplosivo e sui cui sviluppi è difficile fare previsioni.
Cercherò, in questo contributo, di inquadrare il fenomeno analizzando alcuni dei problemi che si presentano a chi effettua una ricerca con alcuni fra gli strumenti più evoluti e potenti oggi disponibili in rete, attraverso una serie di esempi relativi a ciò che si sta facendo per razionalizzare tutta la materia e infine riferendo di alcuni tra i progetti più significativi a riguardo.
Adeguato spazio darò in particolare ai progetti condotti da OCLC — autonomamente o in collaborazione con altri enti — sui temi della descrizione delle risorse: si tratta certamente di progetti tra i più importanti in atto e comunque di particolare interesse per il mondo bibliotecario.
Non è certo questa la prima volta che viene affrontata un'analisi di questo tipo,[1] ma — considerata la velocità con cui tutto si evolve e cambia su Internet — qualsiasi ricerca precedentemente condotta, anche se solo due o tre anni fa risulta incredibilmente datata. Si riportano comunque in nota le indicazioni di precedenti contributi usciti a riguardo e ai quali ho fatto riferimento.[2]

Crescita di Internet

Le cifre indicate nella figura 1 fanno riferimento solo alla crescita di pagine web e limitatamente al biennio del boom 1993-'94.
Oggi la cifra complessiva sembra aggirarsi attorno alle varie decine di migliaia, con un incremento giornaliero stimato intorno ai 50/100 siti.
Si tratta di un fenomeno equiparabile alla produzione editoriale annua di una nazione come la Germania o la Gran Bretagna, con la diffe-

Figura 1 - Crescita di Internet



Fonte: McGraw-Hill)

renza che la fonte di produzione in questo caso può essere anche del tutto individuale e con la differenza che la sua esistenza è legata ad una localizzazione che può essere assolutamente volatile.

Motori, metamotori ed altro...

Parallelamente alla crescita delle fonti primarie (i siti Internet) si è sviluppata quella delle fonti secondarie, vale a dire quei servizi di ricerca in rete o quelle pagine di riferimento che permettono di muoversi all'interno della rete con cognizione di causa. Si parla oggi di motori, metamotori, indici di directory, e così via.
C'è chi ne indica il numero in circa 200, chi meno, indicizzandoli a sua volta in directory di motori e di metamotori.[3] In realtà sono molti di più, se si considerano tutti quelli che hanno un ambito tematico o geografico — in quanto elencano siti che si occupano solo di determinate materie o che riguardano risorse originarie di un determinato paese.
Bisogna poi aggiungere quelli che si definiscono solo come "Bollettini di novità" e che svolgono una funzione pratica molto utile. Sono infatti indici a frequente aggiornamento, con una rotazione delle informazioni che generalmente non supera le due settimane. Molti meno sono quelli a carattere generale; per questa ricerca ne ho censiti 60 su oltre 90 visitati.

Aggiornare Internet

Come avviene l'aggiornamento di questi indici e qual è la loro effettiva estensione e copertura? A parte alcuni, che giustamente rientrano tra i più famosi ed utilizzati, quali Altavista, Yahoo, Webcrawler, Inktomi, Magellan e altri, la stragrande

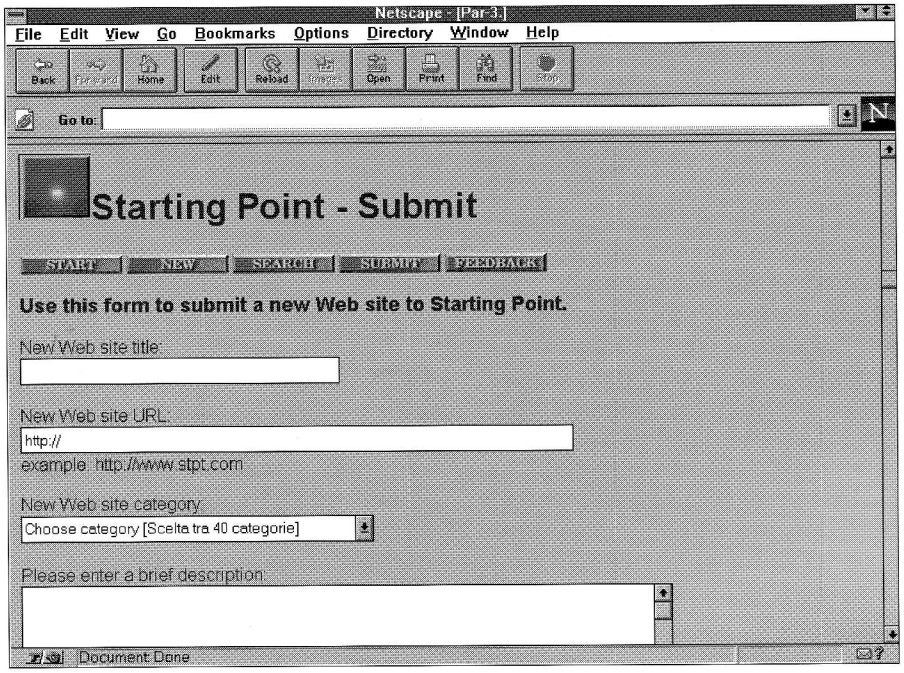
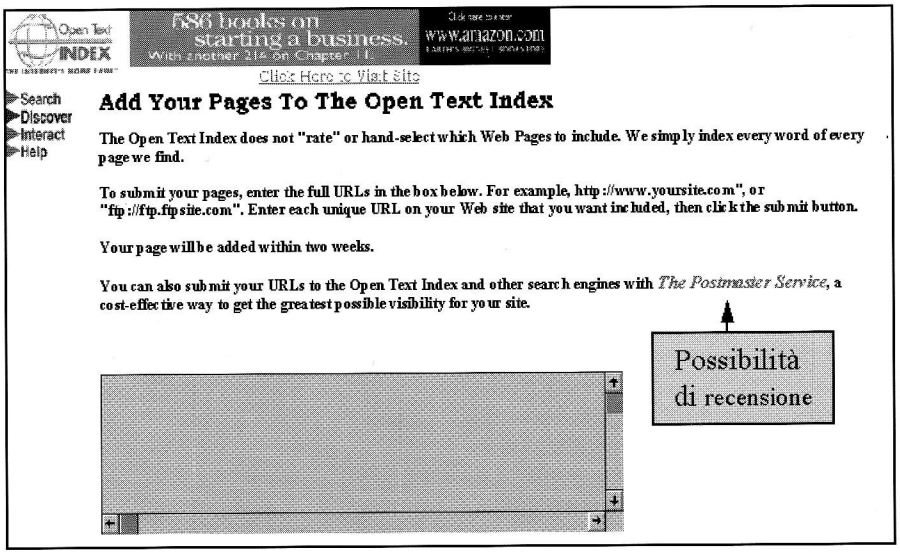


Figura 2 - La maschera di registrazione di un sito su Starting Point

maggioranza usufruisce di aggiornamento volontario, permettendo — anzi stimolando — il casuale visitatore ad aggiungere i dati della propria pagina web o di qualcuna conosciuta. Tutti catalogatori dunque, anche se le modalità possono essere molto differenti da un caso all'altro.
I dati richiesti, come si può vedere su Starting Point[4] (Fig. 2), sono limitati generalmente a pochi campi e la descrizione del sito è lasciata alla serietà dell'autore, oltre che limitata dallo spazio disponibile.
Un'altra fonte di aggiornamento sono gli stessi "Internet provider" commerciali che hanno tutto l'interesse a ospitare pagine web a pagamento, promettendone l'automatica indicizzazione sui principali motori di ricerca.
Nel caso di Open Text[5] (Fig. 3), l'interessato può inserire nel box alcuni dati descrittivi anche generici, oppure chiedere ad un "recen- ➤

Figura 3 - Ancora più ridotta su Open Text



sore" (*Postmaster*) di recensire e verificare il sito. Ovviamente a pagamento!
Va anche detto che i principali motori di ricerca provvedono automaticamente ai propri aggiornamenti, scandagliando la rete alla ricerca di novità, indicizzandole secondo propri criteri, spesso in modo automatico, e recuperando parti iniziali della descrizione originaria.
Un elemento altamente opinabile resta sempre quello relativo alla quantità di informazioni realmente indicizzate e disponibili. Come si

Tabella 1

Webcrawler	145.166 servers
WWWW Worm	3 milioni di url
Open Text	10 miliardi di parole
LYCOS	91% dei siti web
Excite	11,5 milioni di pagine
Altavista	30 milioni di pagine
Inktomi	documents
Harvest	objects/pages
NetFirst	resources

vede dagli esempi[6] elencati nella tabella 1, le unità di misura sono piuttosto difformi anche se forse, al di là della terminologia usata, si tende ad indicare cose analoghe: "document" e "page", URL, siti e "resource" e così via. Non è dato poi conoscere con quale criterio siano scelte (se lo sono) le fonti indicizzate. Ma questo è un problema al quale arriveremo più tardi.
Sia nell'uno che nell'altro caso, molti non dichiarano niente e forse è meglio. Alcuni motori propongono ricerche sulle proprie selezioni attraverso delle raccolte di "review" (recensioni).
Gli autori del motore Excite a tale riguardo sono molti chiari dichiarando di indicizzare effettivamente l'intero contenuto delle pagine web; a differenza di altri, inoltre, non calcolano tra le pagine indicizzate i vari "legami" ad altre pagine. A proposito di chi si comporta diversamente, affermano:

È come se si dichiarasse che una enciclopedia è di 24 volumi di cui 3 di testo effettivo e 21 di "vedi anche" e vari altri rinvii. Crediamo che il modo più onesto per misurare le dimensioni sia quello di fare riferimento alle pagine a testo pieno.

Criteri di scelta

Anche sui criteri di scelta dei siti censiti, ci si mantiene generalmente nel vago. Chiaramente la preferenza va alla quantità delle informazioni disponibili piuttosto che alla qualità. Uniche eccezioni significative: Cyberstacks[7] della Library of Congress e NetFirst di OCLC, che dichiarano esplicitamente di rifarsi a principi rigorosamente biblioteconomici. Cyberstacks è un indice di directory che fa capo alla Library of Congress e che fra l'altro utilizza la classificazione della LC per il proprio ordinamento. Per quanto riguarda la scelta delle fonti, Cyberstacks aderisce ai principi dettati dall'ALA Reference collection development and evaluation committee nel 1992.

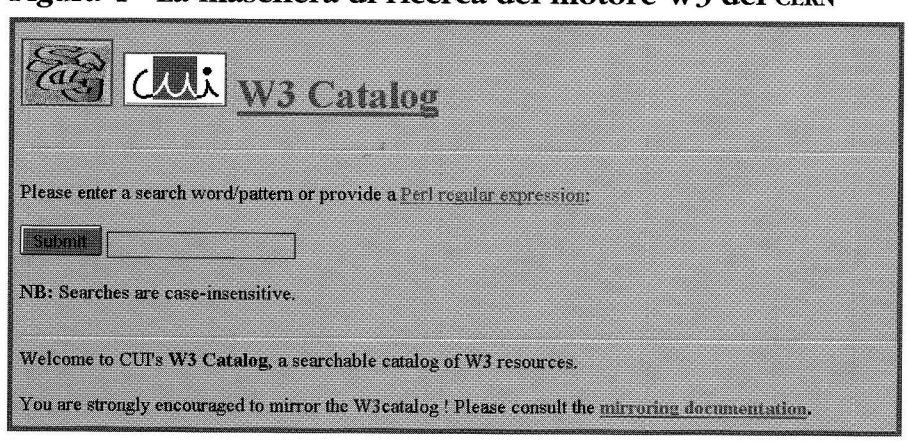
In particolare i criteri indicati da Cyberstacks sono:
- autorevolezza della fonte;
- accuratezza dell'informazione;
- chiarezza della presentazione;
- unicità nella raccolta;
- aggiornamento;
- recensioni favorevoli;
- esigenze della comunità (la propria utenza).

NetFirst, oltre a quelli citati, definisce i propri criteri anche in base alla tipologia della fonte.[8]
McKiernan, curatore di Cyberstacks, indica con molta chiarezza quali siano le differenze tra i due sistemi.[9]

La ricerca

Non sempre chi predispone il "modulo di ricerca" pensa ai suoi potenziali interlocutori. Accade così che anche per l'interfaccia di ricerca ci si imbatta nella più sfrenata varietà e nella mancanza di qualunque standard. Talvolta l'interfaccia è di una semplicità tale da risultare insidiosa.
Con il caso in esempio (Fig. 4) eravamo[10] al limite della semplicità e della complicatezza al tempo stesso. W3 Catalog, un sito sviluppato presso il CERN di Ginevra, proponeva sia una ricerca per termini word/pattern (e sfido chiunque a capire al volo tale espressione) sia una ricerca tramite "Perl regular expression". Il Perl è un linguaggio di programmazione per macchine UNIX, che permette di fare evidentemente ricerche estremamente sofisticate. Di sicuro non è di domi-

Figura 4 - La maschera di ricerca del motore W3 del CERN



nio comune. Non era dato sapere se si potessero utilizzare operatori booleani, troncamenti e altro; non si disponeva di un help ma semplicemente si informava che le maiuscole e le minuscole non venivano tenute in conto.
Ma W3, in realtà, era all'origine di una iniziativa interessante per la standardizzazione della ricerca denominata *mirroring*. In collaborazione con altre istituzioni era stato messo a punto uno strumento sia per condividere risorse in una modalità standard, sia per proporre un "modulo di ricerca" anch'esso standard.
Una specie di comune linguaggio, un modulo di *query*, e dati concatenati e organizzati in modo omogeneo.
Non sempre ci si imbatte in una sinteticità così esasperata. In realtà la maggior parte dei motori dispone di:
- help in linea più o meno dettagliati;
- FAQ (Frequent Asked Questions);
- delimitatori espliciti per tipo di risorsa, paese, tipo di sito (gov, edu, com);
- interfaccia differenziata per la ricerca avanzata.

Tra le altre opzioni abbastanza frequenti vanno menzionate:
- la possibilità di definire l'output (breve/medio/ lungo);
- la possibilità di indicare la fonte su cui indirizzare la ricerca;
- la possibilità di scegliere il numero di risposte da visualizzare.

Permane sempre un certo caos terminologico. Se, infatti, dal punto di vista delle potenzialità di ricerca tutti più o meno ammettono l'uso degli operatori booleani, è anche vero che tutti si sforzano di chiamarli nei più svariati modi.
Si trovano indifferentemente OR e ALL, AND e ANY, NEAR, PROXY e ADJ (per adiacente). L'unico operatore immutabile resta il NOT, quando disponibile.
Non è raro tra l'altro imbattersi in opzioni incomprensibili. Lycos proponeva fino a qualche tempo fa tre formati di visualizzazione: TEXT, TEXT & IMAGE e TERSE TEXT. Chissà cosa intendeva?
L'utilizzo di un certo gergo di tipica impronta anglofona, mette in evidenza un altro dato: la carenza di interfacce multilingue. Nel corso della ricognizione ne ho individuate non più di 5 o 6, la maggioranza delle quali non supera le due/tre lingue. Per Internet, che aspira all'universalità, ci pare francamente una contraddizione, che del resto pare toccare anche la sensibilità dei curatori di alcuni tra i motori più importanti e conosciuti: Altavista ha di recente aperto una opzione europea multilingue e la stessa cosa ha fatto anche Infoseek.

I risultati della ricerca

E finalmente vediamo come vengono presentati i risultati delle ricerche. Col passare del tempo si ha la sensazione che sempre di più i creatori e gestori di questi strumenti si pongano il problema di ottimizzare la presentazione dei risultati aggiungendovi una serie di informazioni ritenute significative, quali:
- la fonte;
- l'indice di rilevanza;
- le occorrenze dei termini cercati;
- la loro evidenziazione nel testo;
- la fonte della eventuale recensione.

Oltre, ai dati comunemente ritenuti indispensabili quali:
- il nome della risorsa;
- l'indirizzo per raggiungerla;
- l'indirizzo di posta del suo curatore (webmaster).

La normalizzazione delle descrizioni e l'individuazione di un formato di riferimento è un problema che da alcuni anni sta tenendo svegli diversi esperti di informazione e bibliotecari in numerose parti del mondo: di alcune delle iniziative già avviate parlerò più avanti. Il problema si sta ponendo a vari livelli anche perché — al di là della disponibilità immediata degli accessi alle risorse — è una descrizione attendibile del loro contenuto che fornisce indicazioni sulla utilità o meno di accedervi, facendo risparmiare tempo prezioso.
La standardizzazione delle descrizioni e la migliore definizione degli accessi, grazie a sistemi di organizzazione (classificazione) delle risorse, sono i due aspetti principali che oggi attendono una risposta.

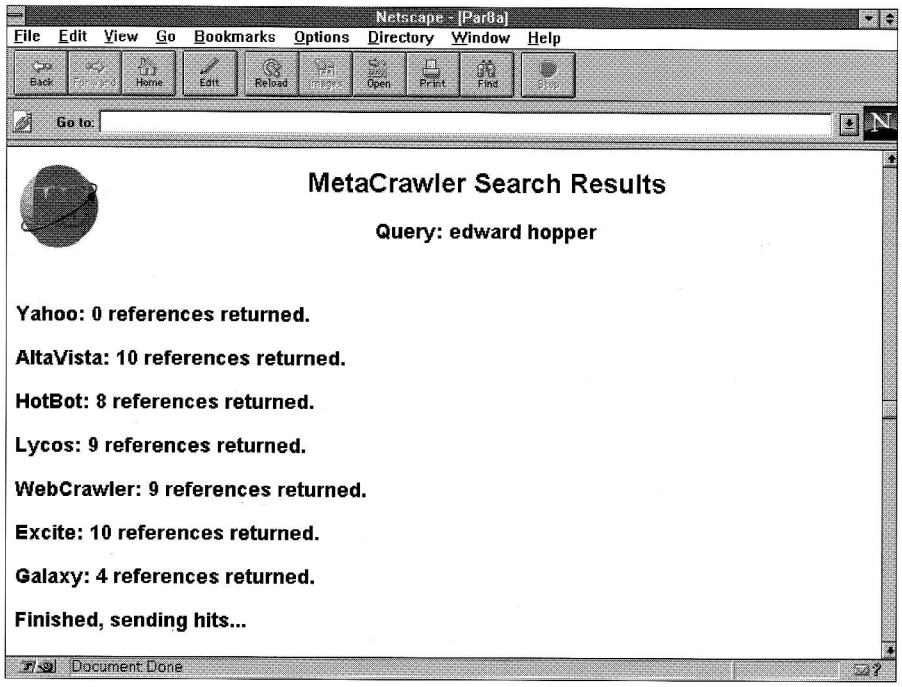
Organizzare Internet

Molti, bibliotecari e non, hanno avuto il primo impatto con Internet grazie alla lettura del libro di Corrado Pettenati e Carla Basili *La biblioteca virtuale*. Al capitolo 5 (il libro è stato scritto nel 1993 e pubblicato nel 1994) i due autori accennano sommariamente alla questione della migliore organizzazione delle risorse Internet, indicandone i due approcci possibili: quello teorico/semantico, di chiara matrice bibliotecaria, secondo il quale "l'accesso alla risorsa viene migliorato imponendo ai dati una struttura" e quello pratico/tecnologico che si realizza grazie ad una "strategia che usa brutalmente la potenza di calcolo e le grandi capacità di trasmissione delle reti per la ricerca su vasti volumi di dati".
Quest'ultimo pare quello finora maggiormente seguito, non foss'altro perché non comporta una dispendiosa attività di indicizzazione.

Metamotori

All'approccio pratico/tecnologico sicuramente si ispirano i cosiddetti metamotori di ricerca, strumenti di intermediazione la cui utilità è strettamente legata al numero di altri motori che essi sono in grado di ➤

Figura 5 - Metacrawler: sommario dei risultati per fonte



interrogare, e alla possibilità o meno di recuperare tutta l'informazione in un'unica lista, possibilmente eliminandone i duplicati. Dai due esempi seguenti credo che la cosa sarà evidente.

Su Metacrawler,[11] effettuata una prima ricerca (Fig. 5) e date le occorrenze sui vari motori il sistema riunisce, eliminando i duplicati, i risultati in un'unica lista. Dalle 50 risposte iniziali se ne ottengono 47 deduplicate (Fig. 6). Vengono mantenute le referenze della fonte in calce.

SavvySearch[12] ha qualche caratteristica in più. Intanto è uno dei pochi strumenti con interfaccia davvero multilingue (ne supporta 19, tra cui il giapponese) e poi va ad interagire con ben 30 motori di ricerca. Permette di selezionare sia il tipo di fonte che il tipo di risultato, la quantità dei dati e la forma della visualizzazione (Fig. 7). Ma la cosa più interessante di Savvysearch è la modalità con cui lavora; interroga infatti in successione laddove ha capito di poter ottenere il numero maggiore di risposte; attraverso una strategia di ricerca fa decidere all'utente se andare avanti o meno.

Si possono inoltre vedere le risposte provenienti dai vari motori anche separatamente prima di richiedere esplicitamente di riunirle in una unica lista.

L'esempio parte dalla stessa ricerca fatta con Metacrawler: nella prima parte (Fig. 8) la lista delle risposte, nella seconda l'indicazione della "strategia" proposta (Fig. 9).

Classificare Internet

Molti indici di risorse Internet si rifanno a schemi di classificazione o di soggettazione di uso corrente, per razionalizzare gli accessi.[13]

Su oltre 40 indici che usano sistemi di classificazione, ben 12 si rifanno alla CDD (Classificazione decimale Dewey), 5 rispettivamente usano la CDU (Classificazione decimale universale) e la LCC (Library of Congress Classification). Infine, 5 indici si ispirano alla LCSH (Library of Congress Subject Heading).

Ancora una volta è McKiernan (autore di Cyberstacks che utilizza la classificazione della LC) a dichiarare le ragioni di una scelta del genere: "Cyberstacks è stato creato come prototipo per dimostrare la

Figura 6 - Metacrawler: i dati deduplicati

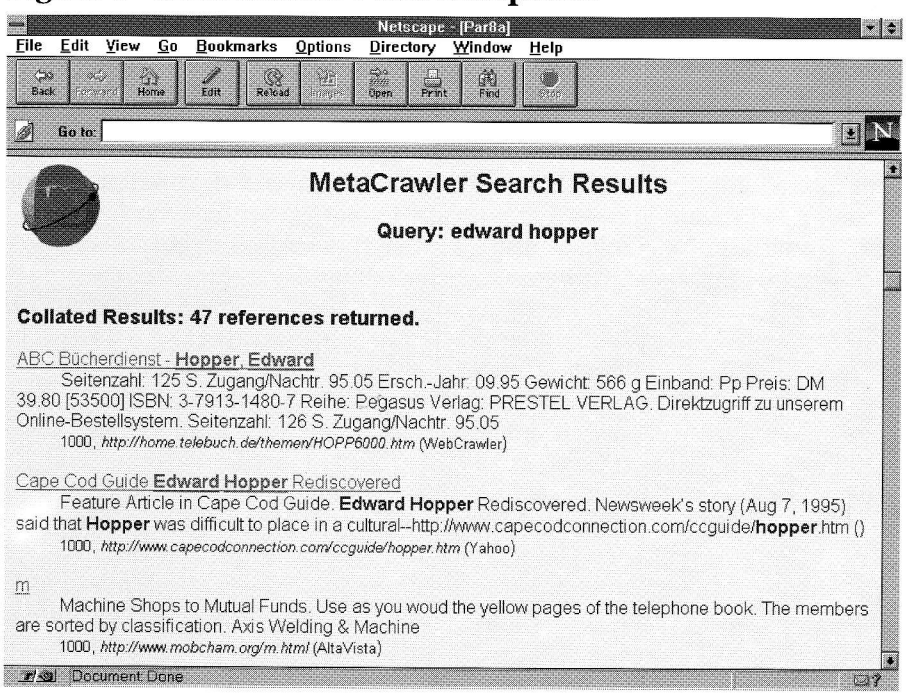


Figura 7 - SavvySearch: il modulo di ricerca

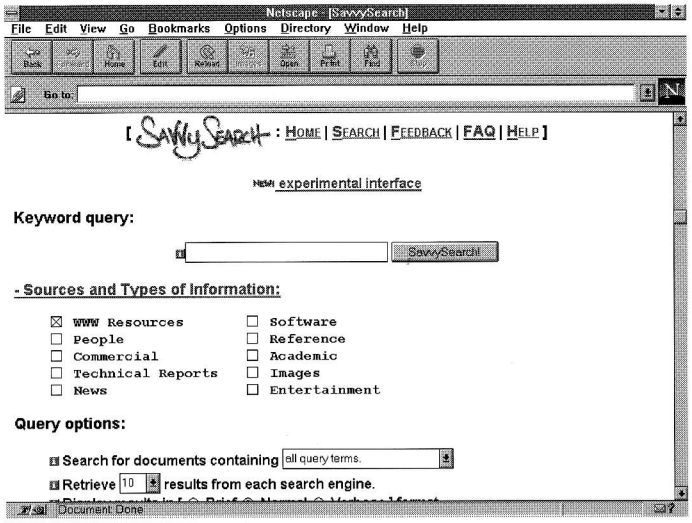


Figura 8 - SavvySearch: la lista delle risposte

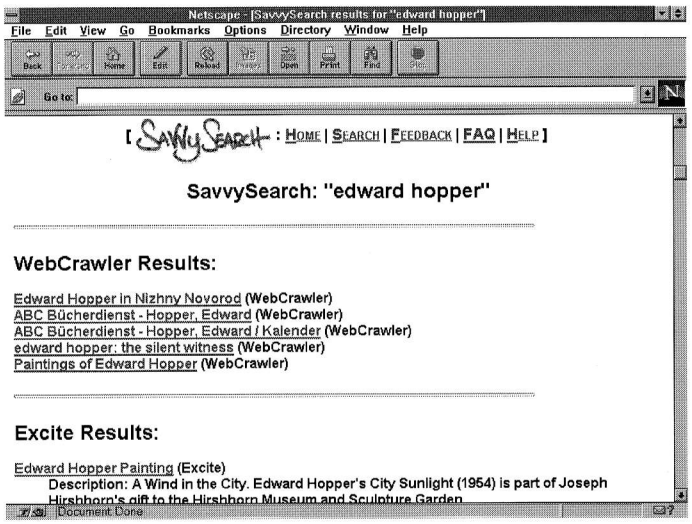
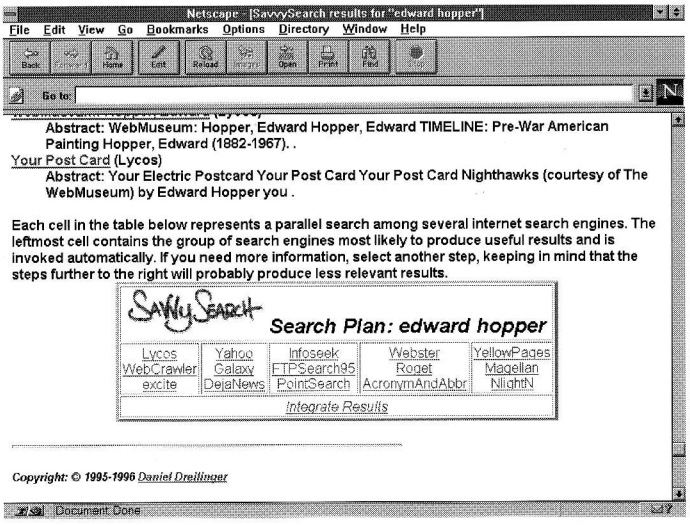


Figura 9 - SavvySearch: la strategia di approfondimento



flessibilità e l'utilità di usare un sistema di classificazione dell'informazione consolidato come uno dei modi per organizzare pagine web ed altre risorse Internet selezionate".
Altri si rifanno alla classificazione Dewey con i medesimi intenti; va detto inoltre che anche per Internet la CDD sta rappresentando un punto di riferimento importante.
Tra i dodici siti finora individuati come utilizzatori della CDD, sicuramente il più articolato ed interessante è CyberDewey;[14] altri sistemi usano la CDD ma in genere non oltre la terza cifra.
Nel caso in esempio (riguardante la classe 780) tratto proprio da CyberDewey, l'utente può navigare dalle 10 classi principali alle 100 suddivisioni dello schema generale fino ad interagire con singole classi, in corrispondenza delle quali sono indicati i siti Internet classificati ed accessibili (Fig. 10 -12, riportate a pagina 24).
Diverso invece l'utilizzo della Dewey che viene fatto su NetFirst, la base dati di risorse Internet creata da OCLC.
In questo caso la CDD può essere utilizzata come un delimitatore della ricerca per termini permettendo, per ora, di scendere al primo livello: nell'esempio (Fig. 13, a pagina 25) la classe 300 (3xx) oppure una suddivisione della medesima, quale la classe 320 "Political sciences" (32x). In futuro sarà possibile selezionare fino alla terza cifra.

Aristotle, Letizia e altri...

Molto si sta lavorando per fornire strumenti avanzati di ricerca, sia per creare accessi semplificati, sia per contribuire all'indicizzazione delle risorse di rete. Un'interessante rassegna è visibile proprio su Internet sotto il nome di Aristotle.[15] Si tratta di oltre 30 progetti o prodotti sponsorizzati da univer- ➤

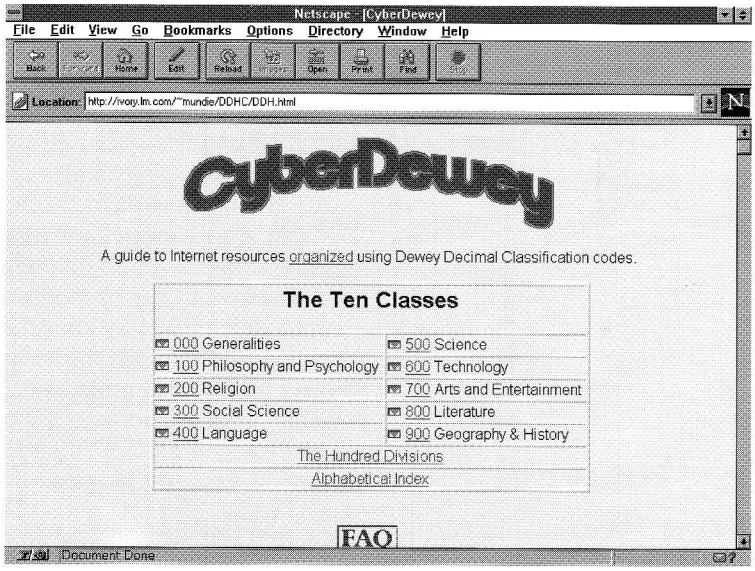
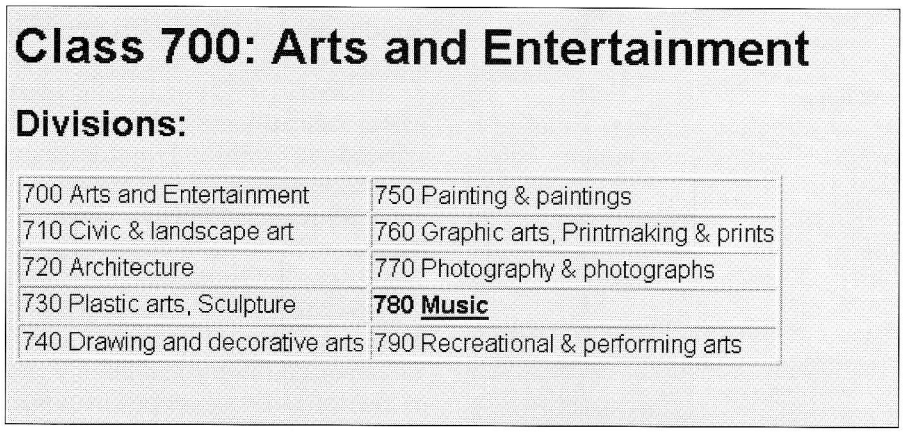
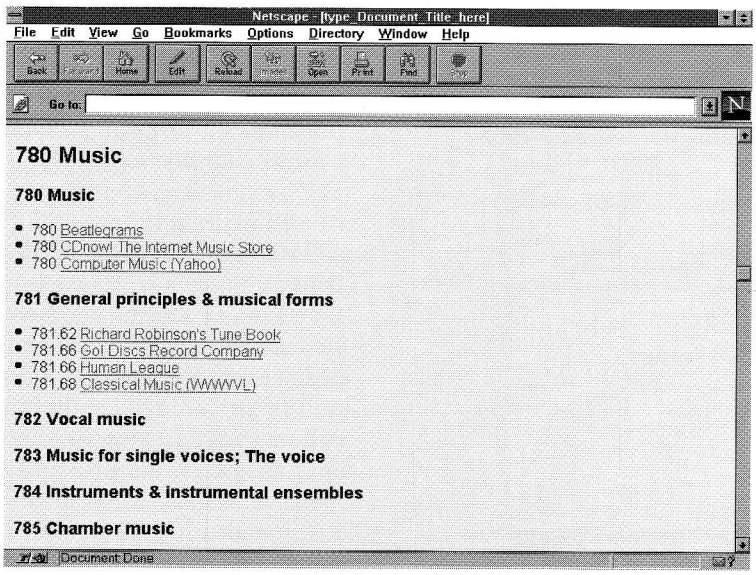


Figura 11 - CyberDewey: navigare nella classe 700





◄ Figura 10 - CyberDewey: la home page

sità, centri di ricerca e aziende. Tra essi ve ne sono numerosi che si propongono di creare delle utility che analizzino ed anticipino, con tecniche mutuate dall'intelligenza artificiale, le esigenze di ricerca dell'utente, fornendogli indicazioni specifiche e lavorando al suo posto nell'individuazione delle fonti che gli potrebbero essere utili.

Un esempio interessante potrebbe essere rappresentato dal progetto Letizia, del quale si puo dire per ora che si tratta di un sw che dovrebbe essere in grado di anticipare i desideri di un ricercatore analizzandone le ricerche già fatte o avviate ed attivando suoi autonomi algoritmi di ricerca. L'utente riceverebbe così indicazioni utili in tempo reale.[16]

Altre utility si propongono invece di scandagliare tutti i messaggi che transitano sulla rete, al fine di raccogliere gli indirizzi dei siti web più citati e costituire così una specie di "Citation index" di Internet, come il progetto PHOAKS (People Helping One Another Know Stuff).[17]

Il progetto "Scorpion"

Tra i progetti di Aristotle troviamo ancora una volta la Classificazione Dewey ed OCLC. Si tratta di Scorpion,[18] un progetto che si propone di combinare la biblioteconomia con le tecniche piu avanzate di reperimento dell'informazione. In pratica Scorpion si propone, con una sofisticata tecnologia di misurazione del testo di un record — nel nostro caso una risorsa Internet — di individuarne le caratteristiche di contenuto; successivamente, confrontando i dati indivi-

◄ Figura 12 - CyberDewey: siti classificati in 780

duati con la griglia di riferimento offerta dalla Classificazione Dewey, Scorpion dovrebbe riuscire a indicizzare automaticamente la risorsa analizzata, collocandola al suo giusto posto.

Descrivere Internet

Ma torniamo al problema principale. Mi pare evidente che le possibilità siano, in termini di funzioni e di potenzialità, ampiamente soddisfatte dalla varietà di soluzioni offerte dai vari motori. Un problema diverso è quello relativo al contenuto dell'informazione. In assenza di qualunque standard descrittivo consolidato, ciascuno si arrangia come può e in diversi casi anche abbastanza bene.
Il tratto comune di quasi tutti gli strumenti esaminati è quello di utilizzare le descrizioni originali della risorsa (una quantità di testo più o meno ampia) senza crearne di proprie. La difformità che ne deriva crea qualche perplessità.
Due o tre esempi scelti fra i tanti chiariranno la questione.
Come potremmo utilizzare la descrizione tratta da Aliweb[19] con le sue circa 40 righe di descrittori? (Fig. 14).
E sarà interessante la Home page della signora Locatelli proposta da Inktomi?[20] (Si veda la figura 15, riportata a p. 26).
Utilizzando descrizioni originali (cioè desunte dal sito) il rischio è anche questo.

Esistono tuttavia lodevoli eccezioni: Webcrawler,[21] pur utilizzando descrizioni desunte risolve il problema con una certa eleganza (Si veda la figura 16 a p. 26).

A questa anarchia descrittiva si oppongono, come alternativa, quei numerosi strumenti di ricerca che si sono dotati di una base dati nella quale includono solo siti re- ➤

Figura 13 - NetFirst: la Dewey per definire la ricerca

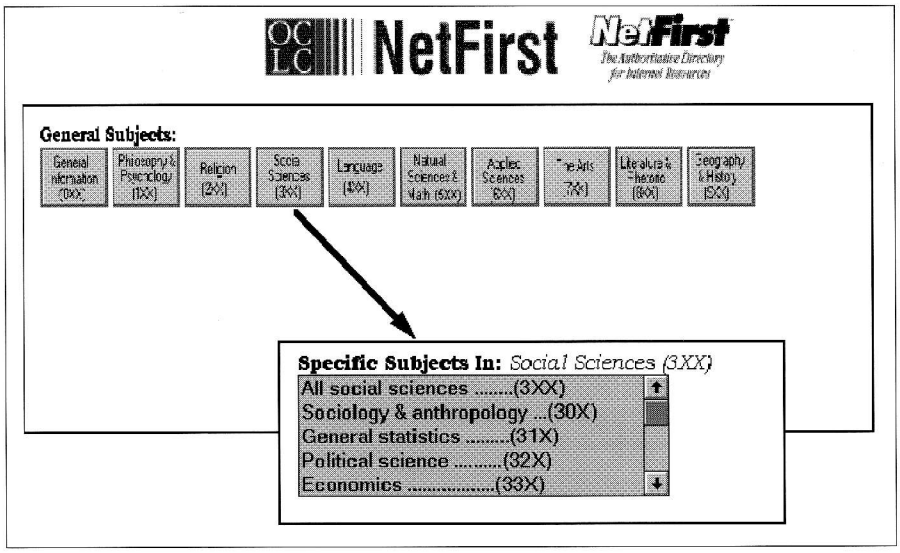


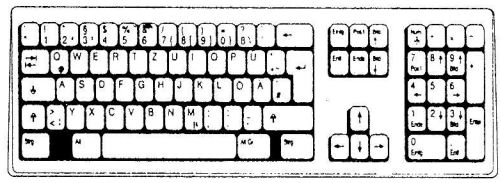
Figura 14 - Un esempio da Aliweb

Search Results for 'libraries"

[50] www.bostonapartments.com

Featuring Classified Ads, Agency Web Pages, Online Client Card, Roommate Matching Service, Articles, Tips and Hot Links to areas of interest in Massachusetts and more!.

Keywords:
Boston Apartments Massachusetts Rentals Real Estate Broker Realtor Agent Pad Flat Apt Room Tenant Landlord Classifieds Roomates Furnish Property Management Government TV Cable Utilities Electricity Gas Radio Boating Sailing Skiing Weather MBTA Colleges Bed and Breakfast Universities Mattresses Futons Management Sales Houses Condominiums Condos Shopping Fanueil Hall Dance Museums Sex Library Limousines Libraries Stadiums Theatre Medical Parks Beaches Jobs Restaurants Movies Cleaning Cleaners Maids Legal Credit Report Agencies Short term Temporary housing house Boston Apartments Massachusetts Rentals Real Estate Broker Realtor Agent Pad Flat Apt Room Tenant Landlord Classifieds Roomates Furnish Property Management Government TV Cable Utilities Electricity Gas Radio Boating Sailing Skiing Weather MTBA Colleges Bed and Breakfast Universities Mattresses Futons Management Sales Houses Condominiums Condos Shopping Fanueil Hall Dance Museums Sex Library Limousines Libraries Stadiums Theatre Medical Parks Beaches Jobs Restaurants Movies Cleaning Cleaners Maids Legal Credit Report Agencies Short term temporary housing house Boston Apartments Massachusetts Rentals Real Estate Broker Realtor Agent Pad Flat Apt Room Tenant Landlord Classifieds Roomates Furnish Property Management Government tv Cable Utilities Electricity Gas Radio Boating Sailing Skiing Weather mtba Colleges Bed and Breakfast Universities Mattresses Futons Management Sales Houses Condominiums Condos Shopping Fanueil Hall Dance Museums Sex Library Limousines Libraries Stadiums Theatre Medical Parks Beaches Jobs Restaurants Movies Cleaning Cleaners Maids Legal Credit Report Agencies Short term temporary housing house



censiti direttamente da loro o da altri. L'utente può così decidere se cercare ovunque oppure mirare su una selezione. Due esempi interessanti tra i tanti possibili, possono considerarsi Magellan e I-Web.[22]

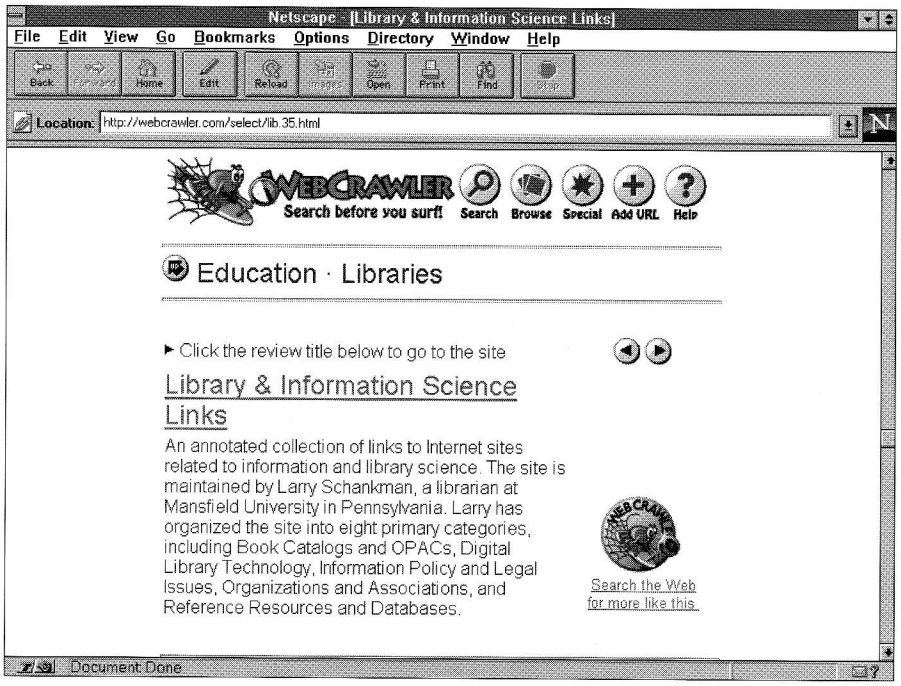
Figura 15 - Un esempio da Inktomi

Return displaying results with

281777 documents satisfied your query: libraries (173941), italy (30416), catalogs (115471).

#5: Score: 881: Angela Locatelli, Home Page
XXXXXXXX—
Relevant words: libraries(6) italy(7) catalogs(5)
http://ccat.sas.upenn.edu:80/rs/al/

Figura 16 - Un esempio da Webcrawler



La descrizione secondo NetFirst

A quanto mi risulta, l'unico indice di risorse Internet che fornisce descrizioni esclusivamente proprie è NetFirst[23] di OCLC, che tra tutti gli strumenti utilizzabili è anche l'unico a pagamento.
Gli elementi caratterizzanti di NetFirst sono:

1) la presentazione in forma di record bibliografico;
2) l'inclusione di un abstract (subject) originale;
3) l'indicazione di soggetti della LC (Library of Congress);
4) l'uso della Classificazione decimale Dewey (Fig. 17).

Netfirst è il frutto di un lavoro che OCLC ha avviato a partire dai primi anni Novanta e che passa attraverso varie iniziative e progetti di studio e approfondimento, il cui obiettivo è quello di arrivare a fissare uno standard descrittivo.

Verso uno standard descrittivo

Sono almeno tre le iniziative degne di essere ricordate.

Figura 17 - NetFirst: una descrizione tipo

Record: 1

DATABASE NO:	96151
TITLE:	HOMEpage OF THE BRAVE: Laurie Anderson
TYPE:	World Wide Web Resources
AUTHOR:	Davies, Jim (Compiler)
PUBLISHER:	Jim Davies
CONTACT:	For Feedback, jimmyd@lanl.gov (Link Internet navigabile)
ACCESS:	(http:)http://www.c3.lanl.gov:8080/cgi/jimmyd/quoter?home (Link Internet navigabile)
DOMAIN:	gov Government
SUMMARY:	Presents information on recording artist, Laurie Anderson. Includes an FAQ section about the artist, a discography of albums, lyrics to several of her songs, films, and videos. Lists books and articles written by and about Laurie Anderson. Notes past performances, reviews of the latest tour, and a tour schedule. Provides lyric interpretations, art and poetry inspired by Anderson, and a term paper about Anderson.
LC SUBJECT:	Anderson, Laurie (Singer). Singers. Popular music.
DEWEY CLASS:	783.0092 781.63

1."Assessing the information on the Internet"
Questo progetto, svoltosi nel biennio 1992-'93, ha costituito l'approccio pionieristico al problema delle descrizioni ed ha avuto come oggetto fondamentale l'individuazione e la quantificazione delle varie tipologie di risorse presenti all'epoca sulla rete. Inoltre ha esaminato se e come il formato MARC potesse essere utilizzato per descrivere tali risorse ed infine ha indicato alcune modifiche da apportare al formato MARC stesso per poter compiutamente descrivere la particolare natura delle risorse virtuali: in particolare l'inserimento del nuovo campo (tag) 856 – Location.

2. "Intercat"
"Intercat" è stato il primo progetto sperimentale di catalogazione di risorse Internet svolto in cooperazione da circa 30 biblioteche statunitensi, utilizzando le AACR2 e il formato USMARC. Si è svolto tra l'ottobre 1994 ed il marzo 1996 e ha prodotto una base di circa 5.700 registrazioni bibliografiche. Tutti i record prodotti sono accessibili su NetFirst, è in elaborazione il report finale.

3. "Metadata Workshop"
Condotto nel 1995 da OCLC in collaborazione con il NCSA (National Center of Supercomputer Applications) ha avuto come scopo quello di focalizzare maggiormente il problema della creazione di descrizioni standard, in funzione del reperimento e dello scambio di dati.
È da questo workshop che è scaturita la proposta di elaborare un metaformato di descrizione delle risorse Internet, conosciuto convenzionalmente con il nome di Dublin Core.

Il metaformato Dublin Core

Dublin Core non si propone come sostituto di altri formati esistenti (per esempio del MARC o di sistemi consolidati di classificazione) ma piuttosto come formato integrativo, appunto un metaformato. Questo per garantire che almeno un certo tipo di dato sia comunque presente in una descrizione relativa ad una risorsa Internet.[24] Senza entrare nel dettaglio dei singoli campi, (vedi tabella 2), mi limito ad osservare che Dublin Core in quanto metaformato prevede l'accoglimento di qualsiasi altro codice standard.[25] Campo per campo, infatti, si può segnalare lo standard che viene adottato evidenziandolo con alcuni elementi convenzionali quali: scheme = per i titoli e role = per gli autori.
Nella tabella 3 si vede come Dublin Core potrebbe trattare dati di diversa provenienza e con diverse caratteristiche relativamente al soggetto, al titolo o all'autore.

Conclusioni

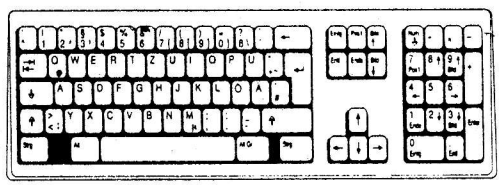
Considerata la velocità con cui ormai tutto si evolve credo di poter dire che per ora il punto a cui si è arrivati sia sostanzialmente quello che ho esposto. Era ovvio che — nel momento in cui ci si è posti seriamente il problema di organizzare informazioni e soprattutto di creare "metaformati" per gestire "metadati" — i bibliotecari, produttori da sempre di "metadati" — dal repertorio inventariale fino al record catalografico — si trovassero nel proprio habitat ideale e si sentissero doverosamente impegnati a fornire un proprio contributo.
Uno studioso di Internet, analizzando l'esigenza di ricorrere ai "metadati" per gestire le risorse virtuali, ha ricordato che: "Visto che Platone e Cartesio assimilavano la conoscenza a 'cibo per la mente', se oggi riteniamo ancora che sia così, allora è venuto il momento di pensare a una dieta". ■ ➤

Tabella 2 - *Dublin Core* Element Description

- SUBJECT: The topic addressed by the work
- TITLE: The name of the object
- AUTHOR: The person(s) primarily responsible for the intellectual content of the object
- PUBLISHER: The agent or agency responsibile for making the object available
- OTHERAGENT: The person(s), such as editors and transcribers, who have made other significant intellectual contributions to the work
- DATE: The date of publication
- OBJECTTYPE: The genre of the object, such as novel, poem, or dictionary
- FORM: The data representation of the object, such as Postscript file or Windows executable file
- IDENTIFIER: String or number used to uniquely identify the object
- RELATION: Relationship to other objects
- SOURCE: Objects, either print or electronic, from which this object is derived, if applicable
- LANGUAGE: Language of the intellectual content
- COVERAGE: The spatial locations and temporal durations characteristic of the object

Tabella 3

- Subject (scheme=LCSH) = UNIX (Computer system)
- Subject (scheme=Dewey Decimal System) = 004.251 Supercomputers–systems design
- Title (scheme=AACR2) = The structure of language
- Author (scheme=USMARC) = 100 1 Doyle, Arthur Conan $c Sir, $d 1859-1930



Note

[1] I due contributi che ho trovato di maggiore interesse sono quelli di: JIAN LIU, *Understanding www search tools*, September 1995, February 1996 (http://www.indiana.edu/~librcsd/search/) e di H. VERNON LEIGHTON, *World wide web indexes: a study*, June 1996. (http://www.winona.msus.edu/services-f/library-f/webind.htm).
Per ulteriori contributi si rimanda a *Understanding and comparing WEB search tools* (http://www.hamline.edu/library/links/comparisons.html).

[2] Questo testo riporta, con qualche aggiornamento, il mio intervento in occasione del V Workshop "Information in Libraries" organizzato dall'Università Cattolica del Sacro Cuore di Milano in collaborazione con la società CENFOR di Genova. Questo contributo, con il titolo *Strumenti per la ricerca ed il recupero dell'informazione su Internet*, completo di tutti i link alle risorse Internet citate, è gentilmente ospitato dal sito della Libreria Burioni di Genova ed è accessibile all'URL http://www.burioni.it/forum.

[3] Due indici interessanti sono All-in-one (http://www.albany.net/allinone/) e quello predisposto dal NCSA (National Centre for Supercomputer Applications) di San Diego in California (http://www.ncsa.uiuc.edu/SDG/Software/Mosaic/MetaIndex.HTM).

[4] Starting Point, http://www.stpt.com/.

[5] Open Text, http://www.opentext.com/.

[6] Webcrawler, http://webcrawler.com/GNN/WebQuery.HTML; WWWW Worm, http://wwww.cs.colorado.edu/wwww; Lycos, http://www.lycos.com/; Excite, http://www.excite.com/; Altavista, http://altavista.digital.com/; Inktomi, http://inktomi.berkeley.edu/query.HTML; Harvest, http://harvest.cs.colorado.edu/; NetFirst, http://www.oclc.org/oclc/netfirst/netfirst.HTM.

[7] Cyberstacks, http://www.public.iastate.edu/~CYBERSTACKS/homepage.HTML.

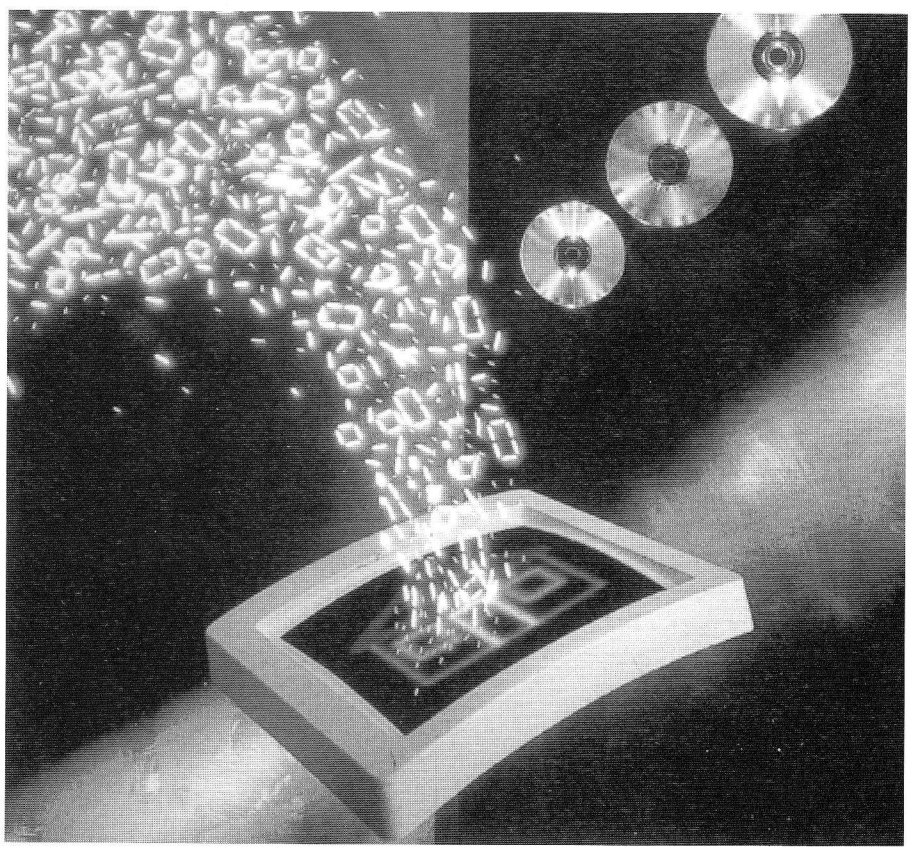
[8] Da *NetFirst collection development policy selection. Principles. Insofar as they apply to online resources, NetFirst's collection principles will follow the American Library Association's "Library Bill of Rights."* (Appendix A.). *NetFirst will include objects in all subject areas.*

- InterCat records
- OCLC sources
  Networks
  Membership lists
  Governance
  Advisory committees
  User suggestions
  Periodicals
  Library-focused
  Internet focused
  Subject-focused, for special areas, for example: Fortune magazine, for list of the "Fortune 1000" to insure their inclusion
- Internet object review sites and object collections
- Library sites
- National, state, provincial and local government sites.

[9] G. MCKIERNAN, *Cyberstacks: selection, description, incorporation and presentation*, "Although NetFirst© does include significant monographic and serial works, files, databases and search services, it does not differentiate between discrete resources and entire web sites in its selection. [...]. NetFirst© does not differentiate between individual resources and more comprehensive or complex collections. CyberStacks(sm) on the other hand, intentionally seeks to identify and describe only discrete resources, be they unique or parts of a larger collection".

[10] La schermata che vedete nella Figura 4 può essere considerata un pezzo di "archeologia" di Internet. W3 infatti è stato dichiarato morto l'8 novembre 1996. Sono interessanti le motivazioni del "suicidio" di cui riportiamo un paragrafo.
"W3 Catalog is Dead! Why has W3 Catalog been stopped?".
"Although W3 Catalog was very popular, it has been made obsolete for a number of technical and practical reasons.

1. Maintenance overhead: W3 Catalog was designed to re-build itself automatically without administrator intervention. Unfortunately there is still a certain amount of work involved in keeping track of the lists that are consulted, making sure that URLS are valid, and answering email requests. Nobody has time or interest for this.
2. Fragility: The weak link in the chain is the chunking. If the format of the



lists changes slightly, the chunking script can break, and the catalog will start to contain "ugly" records (i.e., records that are several pages long, or records that don't form complete or valid HTML paragraphs). Writing a more robust chunker is a non-trivial task.

3. Implementation overhead: Htgrep was designed to be a simple, general-purpose back-end to ad hoc search engines. It is cheap (free!), portable (Perl), and relatively simple to install (well, there is a FAQ). Unfortunately Htgrep does not lend itself very well to applications like W3 Catalog because of its extreme simplicity: For each request, a new process is created to run an instance of the htgrep CGI script. This, of course starts a copy of the Perl compiler and reads and compiles htgrep for each request. For each request, the entire W3 Catalog database is read and scanned. This overhead is fine for small catalogs or for relatively infrequently accessed search engines. But the W3 Catalog database is several megabytes, and, at peak times, several requests are received every second. This quickly brings the CUI server to its knees.
4. Competition: Frankly, much better, faster, and more comprehensive search engines are now available. Alta Vista and HotBot are two popular search engines whose contents are generated by periodically scanning the entire www. Their implementations are certainly not based on CGI, and are highly optimized".

[11] Metacrawler, http://metacrawler.cs.washington.edu:8080/.

[12] Savvysearch, http://www.cs.colostate.edu/~dreiling/smartform.HTML.

[13] Per un aggiornamento sulla lista vedi: G. MCKIERNAN, *Beyond bookmarks: schemes for organizing the web*, "... a clearinghouse of World Wide Web sites that have applied or adopted standard classification schemes or controlled vocabularies to organize or provide enhanced access to Internet resources". http://www.public.iastate.edu/~CYBEZRSTACKS/CTW.HTM.

[14] CyberDewey, http://ivory.lm.com/~mundie/DDHC/DDH.HTML.

[15] Aristotle, http://www.public.iastate.edu/~CYBEZRSTACKS/Aristotle.HTM.

[16] H. LIEBERMAN, *An automated channel-surfing interface agent for the Web*. Presented at the artificial intelligence-based tools to help W3 users workshop, May 6, 1996, the Fifth International World Wide Web Conference, May 6-10, 1996, Paris, France. (http://www.info.unicaen.fr/~serge/3wia/workshop/ papers/paper29.HTML) e *Letizia: an agent that assists Web browsing*. Paper presented at the International Joint Conference on Artificial Intelligence, August 20-25, 1995, Montreal, Canada. (http://lieber.www.media.mit.edu/people/ lieber/Lieberary/Letizia/Letizia.HTML).
L'interesse per questi strumenti sofisticati si sta estendendo anche a livello commerciale. Mi riferisco ai cosiddetti servizi di "Software Agent" già disponibili sulla rete, quali Firefly (http://www.ffly.com) e altri.

[17] PHOAKS, http://weblab.research.att.com/phoaks/.

[18] Scorpion, http://purl.oclc.org/scorpion.

[19] Aliweb, http://web.nexor.co.uk/public/aliweb/aliweb.HTML.

[20] Inktomi, http://inktomi.berkeley.edu/query.

HTML.

[21] Webcrawler, http://webcrawler.com/GNN/WebQuery.HTML.

[22] Magellan, http://www.mckinley.com/; I-Web: http://sparta.lcs.mit.edu/iweb.

[23] NetFirst, http://www.oclc.org/oclc/netfirst/netfirst.HTM.

[24] Per dovere di completezza mi sembra giusto citare anche un'altro interessante tentativo di razionalizzazione delle descrizioni, effettuato dalla McKinley, creatrice del motore Harvest. Si tratta di Summary Object Interchange Format (SOIF), concepito però più come formato di scambio dati. http://harvest.transarc.com/afs/transarc.com/public/trg/Harvest/technical.HTML.

[25] Una descrizione ampia e dettagliata del formato e dei lavori che lo hanno preceduto e che stanno proseguendo è visibile sulle pagine dedicate al Metadata Workshop e ospitate dal Web di OCLC. http://www.oclc.org:5046/oclc/research/conferences/metadata/.

[26] L. FLORIDI, *The Internet: which future for organised knowledge?*, "The electronic library", 14 (1996), 1.