

Informazione biomedica su internet e motori di ricerca. Risultati di un trial di un anno

Salvatore Corrao¹, Francesco Leone², Sabrina Arnone¹

Riassunto. Internet rappresenta oggi un importante mezzo di comunicazione e un enorme contenitore di informazioni che può risultare di grande utilità nella ricerca e nel recupero dell'informazione biomedica. Tuttavia, questa eccessiva disponibilità rappresenta nel contempo un limite alla fruizione di informazione utile. I motori ed i metamotori di ricerca rappresentano lo strumento principe che permette una rapida ricerca dell'informazione, anche in campo biomedico, contenuta sul web ma, a nostra conoscenza, non sono disponibili dati su quanto essi siano affidabili, soprattutto da un punto di vista della riproducibilità dei risultati. Per tali motivi, lo scopo del nostro studio è stato quello di verificare la riproducibilità di una ricerca per parole chiave ("pediatric" o "evidence"), interrogando nel novembre 2001 e dopo un anno, 9 motori di ricerca internazionali ed un metamatore. Abbiamo analizzato le prime 20 citazioni, in ordine di rilevazione, per ogni motore di ricerca, e suddiviso i risultati in base all'estensione dei siti. Abbiamo, quindi, confrontato la ricerca effettuata all'inizio dello studio con quella effettuata dopo un anno e considerato come criterio di affidabilità la riproposizione degli stessi siti a distanza di un anno. I nostri risultati (descritti nel testo) sottolineano l'estrema dinamicità dell'informazione sulla rete e raccomandano la massima cautela nell'utilizzo dei motori di ricerca generici per il recupero dell'informazione biomedica sul web. Tuttavia, pensiamo che questi motori di ricerca possano rappresentare un ottimo strumento per individuare siti istituzionali e fornire risultati utili per focalizzare meglio una ricerca, da effettuare in seguito con mezzi orientati esclusivamente all'informazione biomedica. Pensiamo, infine, che questo studio dia un contributo ad un approccio più consapevole all'universo dell'informazione biomedica su internet.

Parole chiave. Informazione biomedica, internet, metamotori e motori di ricerca.

Summary. *Biomedical information on the internet using search engines. A one-year trial.*

The internet is a communication medium and content distributor that provide information in the general sense but it could be of great utility regarding as the search and retrieval of biomedical information. Search engines represent a great deal to rapidly find information on the net. However, we do not know whether general search engines and meta-search ones are reliable in order to find useful and validated biomedical information. The aim of our study was to verify the reproducibility of a search by key-words (pediatric or evidence) using 9 international search engines and 1 meta-search engine at the baseline and after a one year period. We analysed the first 20 citations as output of each searching. We evaluated the formal quality of Web-sites and their domain extensions. Moreover, we compared the output of each search at the start of this study and after a one year period and we considered as a criterion of reliability the number of Web-sites cited again. We found some interesting results that are reported throughout the text. Our findings point out an extreme dynamicity of the information on the Web and, for this reason, we advice a great caution when someone want to use search and meta-search engines as a tool for searching and retrieve reliable biomedical information. On the other hand, some search and meta-search engines could be very useful as a first step searching for defining better a search and, moreover, for finding institutional Web-sites too. This paper allows to know a more conscious approach to the internet biomedical information universe.

Key words. Biomedical information, internet, meta-search engines, search engines.

¹ Unità Operativa di Metodologia Clinica ad indirizzo epidemiologico-statistico, ² Unità Operativa VI Pediatria per le Emergenze; Azienda di Rilievo Nazionale ad Alta Specializzazione Ospedali Civico e Benfratelli, G. Di Cristina, M. Ascoli, Palermo.

Pervenuto il 2 aprile 2003.

Introduzione

Le pagine web sono create da una grande varietà di organizzazioni ed il tipo di organizzazione può rappresentare una indicazione di quanto affidabile ed obiettiva sia l'informazione contenuta. Infatti, un semplice modo per riconoscere la fonte e la qualità della informazione può essere rappresentato dall'analisi delle estensioni dei siti pubblicati sul web (tabella 1). Per esempio, la Walden University¹ propone quale criterio per una ricerca efficace dei siti web in campo biomedico proprio quello di considerare i siti con estensione .com e quelli con estensione di tipo nazionale meno attendibili rispetto a quelli con estensione .edu e .gov, in primo luogo, e in minor misura, .org.

Comunque, si deve tener presente che l'accesso ai diversi siti internet, e questo vale anche per i siti la cui informazione è di natura biomedica, avviene spesso (fino all'80% degli accessi, secondo varie fonti di rilevanza)²⁻⁵ tramite i classici i motori di ricerca e portali di interesse generale.

Per tali motivi, abbiamo voluto valutare la riproducibilità di una ricerca per parole chiave, utilizzando i principali motori di ricerca in periodi differenti (a distanza di un anno).

Tabella 1. - Estensioni dei siti web e loro significato.

.com	è in genere un sito commerciale, può essere una sorgente di informazioni in grado di fornire informazioni correnti o attuali oppure può trattarsi di siti sponsorizzati da una impresa commerciale, spesso interessata soltanto a vendere
.edu	sito sponsorizzato da una istituzione educativa (ospedaliera o universitaria), di solito attentamente monitorizzato e, quindi, in genere affidabile e di qualità.
.gov	sito governativo, spesso una buona sorgente di dati o informazioni utili, di solito rigidamente controllato per la sua stessa origine istituzionale
.mil	sito militare, l'accesso è sovente di tipo riservato
.net	sito di rete costruito e messo a disposizione dei suoi sottoscrittori con poca o nessuna supervisione; coloro che pubblicano queste pagine possono anche essere affiliati ad istituzioni scolastiche o accademiche
.org	sito di una organizzazione pubblica o privata o non profit; alcuni gruppi possono risentire dell'influenza della pubblica opinione ma altri sono buone sorgenti di informazioni
.paese*	sito che individua la nazione di registrazione; questo tipo di estensione non consente di distinguere la tipologia del sito stesso, che può essere di qualsiasi natura; l'unica informazione estrapolabile è la lingua prevalentemente utilizzata
~ seguito dal nome personale	indica una pagina di tipo personale; alcune istituzioni educazionali consentono ai singoli la pubblicazione di pagine personali; scarso o assente il controllo dei contenuti

* Il suffisso .paese include tutte le estensioni a carattere nazionale (ad es. uk, .ca, .it, etc)

Metodi

Abbiamo interrogato nove motori di ricerca internazionali fra i più diffusi ed un metamatore di ricerca (tabella 2), utilizzando due parole chiave: "pediatric" ed "evidence" con i seguenti criteri.

– Ricerca tutte le parole selezionando il criterio "any words" (quando è risultata disponibile solo la ricerca semplice) oppure l'operatore booleano "OR" (quando è risultata disponibile la ricerca avanzata);

– ricerca in tutto il web.

Tabella 2 - Elenco dei motori di ricerca e modalità di ricerca utilizzate.

Motori di ricerca	Modalità di ricerca
Altavista	semplice
Google	semplice
Goto	semplice
Hotbot	avanzata
Yahoo	semplice
Lycos	semplice
Msn	semplice
Supereva	semplice
Excite*	semplice
Metamatore di ricerca Webcrawler	semplice

* Nel 2002 Excite è diventato un metamatore

Abbiamo ritenuto che, oltre alla parola chiave "evidence" (il termine più specifico che caratterizza il concetto di Evidence Based Medicine), poteva risultare interessante l'associazione con una parola chiave più specialistica come "pediatric".

Abbiamo considerato i primi 20 risultati, in ordine di citazione, per ogni motore di ricerca e suddiviso i risultati in base all'estensione dei siti.

Abbiamo effettuato questa ricerca nel mese di novembre 2001 e l'abbiamo ripetuta ad un anno esatto di distanza, nel mese di novembre 2002.

Abbiamo quindi valutato la riproposizione degli stessi siti a distanza di un anno.

Risultati

Sulla base del criterio di riproposizione sopra citato, abbiamo stilato una graduatoria dei motori di ricerca utilizzati: la più alta percentuale di riproposizione è stata rilevata per soli tre motori di ricerca (Google, Goto e Yahoo) e la più bassa, pari allo 0%, solo per un motore di ricerca: Lycos (tabella 3 alla pagina 24).

Sembra rilevante segnalare il fatto che, nonostante la bassa percentuale di riproposizione, per ben 8 dei motori di ricerca considerati (Supereva, Msn, Janas, Yahoo, Altavista, Hotbot, Goto e Google) il sito proposto per primo nel 2001 è lo stesso proposto per primo nel 2002, e ancora che si tratta dello stesso sito per tutti e 8 i motori di ricerca (<http://depts.washington.edu/pedebm/>). Inoltre, quest'ultimo non è un sito commerciale, anzi l'estensione presuppone a priori che il sito sia di tipo informativo-formativo; ciò è confermato dall'analisi del contenuto, che sembra affrontare compiutamente l'argomento "pediatria basata sulle evidenze". In nessun altro caso uno dei siti riproposti nel 2002 occupa la stessa posizione che occupava nel 2001 (figura 1 alla pagina seguente).

Tabella 3. - *Elenco dei motori di ricerca ordinati per percentuale di riproposizione dei siti.*

Motori di ricerca	Siti riproposti	% di riproposizione*
Goto	6	30
Google	5	25
Yahoo	5	25
Supereva	4	20
Hotbot	3	15
Msn	3	15
Webcrawler	3	15
Altavista	2	10
Lycos	0	0

* Sul totale dei primi 20 siti della lista proposta da ogni motore di ricerca rispetto a quella dell'anno precedente.

N.B. Excite non è riportato perché è diventato nel 2002 un metamatore di ricerca.

Abbiamo inoltre effettuato una valutazione comparativa dei tipi di estensione ritrovati. Non abbiamo riscontrato significative differenze fra il 2001 e il 2002 per quanto riguarda il numero dei siti con suffisso .edu e .gov. Al contrario, abbiamo riscontrato una riduzione dei siti con suffisso .org, e dei siti con estensione .paese, e un incremento dei siti con estensione .net e .com (tabella 4, alla pagina a fronte).

Nella tabella 4 non sono riportati i dati relativi ad uno dei 9 motori di ricerca indicati nella sezione "metodi". Infatti, Excite utilizzato nella ricerca del 2001 come motore di ricerca, è diventato nel 2002 un metamatore di ricerca entrando nell'orbita di Infospace⁶. I siti web derivati dalla sua ricerca sono stati, quindi, esclusi dall'analisi finale dei risultati. Abbiamo, comunque, pensato di confrontare la ricerca condotta nel 2002 su Excite con quella condotta, sempre nel 2002, sull'altro metamatore di ricerca, Webcrawler, ottenendo una quasi completa concordanza dei risultati (figura 2, a pagina seguente).

Nella tabella 4, l'estensione .paese include tutti i siti web aventi una estensione relativa al paese di registrazione.

Discussione

I motori di ricerca, nati con lo scopo di ricercare velocemente pagine web pubblicate su internet, si sono evoluti in tre direzioni definite, ma dai confini comunque sfumati.

I motori di ricerca propriamente detti (search engine): scandagliano continuamente la rete, analizzano le pagine di ogni sito, le indicizzano e creano i relativi riferimenti nei loro database. Questo lavoro viene eseguito automaticamente da robot (chiamati crawler o spider), cioè da programmi che esplorano il web senza intervento dell'uomo.

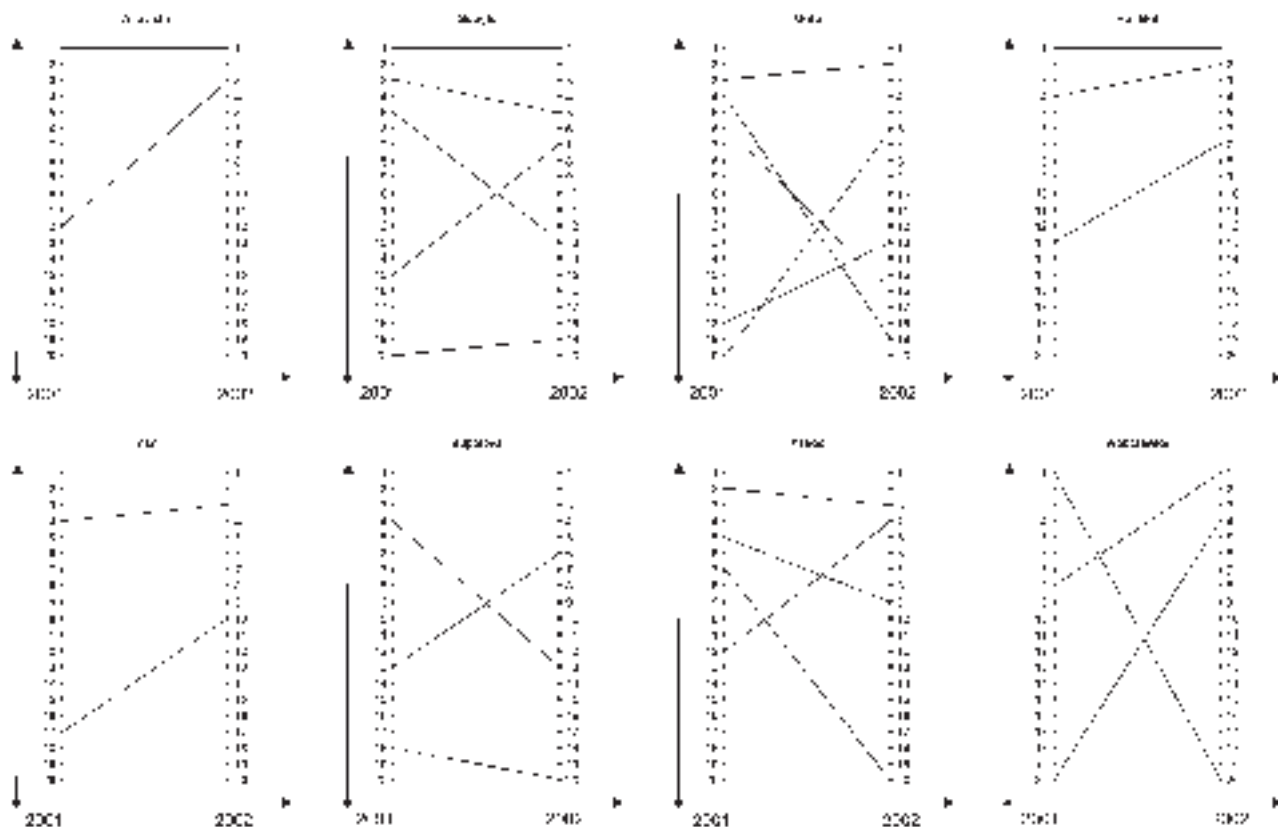


Figura 1. Cambiamento della posizione dei siti tra il 2001 e il 2002 per ogni motore di ricerca. Le barre indicano, a partire dalla posizione del 2001, l'eventuale riproposizione ed il cambiamento di posizione del sito nel 2002.

Tabella 4 - Valutazione comparativa (2002 versus 2001) del numero di siti ritrovati suddivisi sulla base dell'estensione.

	.edu		.gov		.org		.net		.com		.paese	
	2001	2002	2001	2002	2001	2002	2001	2002	2001	2002	2001	2002
Motori di ricerca												
Altavista	9	3	0	4	5	2	0	2	5	9	1	0
Google	8	8	0	0	3	3	1	0	4	6	4	3
Goto	5	8	0	0	3	1	0	1	7	7	5	3
Hotbot	5	7	0	1	5	3	0	1	6	6	4	2
Yahoo	6	7	0	1	3	3	0	0	7	6	4	3
Lycos	9	4	2	3	1	4	1	1	3	7	4	1
Msn	6	7	0	0	5	2	0	1	8	9	1	1
Supereva	7	8	0	0	3	2	0	0	6	7	4	3
Metamotore di ricerca												
Webcrawler	7	8	1	0	5	1	0	1	7	10	0	0
Totale di citazioni	62	60	3	9	33	21	2	7	53	67	27	16
Valore mediano	7	7	0	0	3	2	0	1	6	7	4	2

* Il suffisso .paese include tutte le estensioni a carattere nazionale (ad es. uk, .ca, .it, etc)
N.B. Excite non è riportato perché è diventato nel 2002 un metamotore di ricerca.

(Per esempio: Scooter per Altavista, Smartcrawl per Hotbot, T.Rex per Lycos, Gulliver per Nothern light, etc). Ciascun programma di ricerca utilizza regole proprie per mettere in ordine (ranking) e presentare gli indirizzi trovati. Lo scopo di questo tipo di motore di ricerca, comunque, è cercare di disporre nei suoi database il maggior numero di pagine alle quali sono associati riferimenti, ma non esprime alcun giudizio qualitativo sui siti.

Nel tempo, i motori di ricerca si sono evoluti e la loro strategia di ricerca è passata dalla semplice indicizzazione delle parole chiave alla valutazione automatica della pertinenza del contenuto dell'intera pagina e del numero dei collegamenti al sito.

I motori di ricerca contenuti nei "portali" rappresentano una ulteriore evoluzione. I portali non forniscono riferimenti a singole pagine, ma ai siti che vengono classificati, all'interno dello stesso portale, in categorie o directory, grazie all'intervento di operatori umani (surfer). Esistono due tipi di portali: generici (come ad esempio Yahoo, che fornisce una agevole esplorazione guidata del web grazie alla suddivisione ad albero della sua directory, oppure Open Directory in cui le categorie di siti vengono create da operatori volontari), e specifici che includono categorie di siti web limitati a particolari settori specializzati come, ad esempio, medicina, ricerca scientifica, affari, finanza, news, etc. Questi motori di ricerca permettono, oltre la ricerca sull'intero web, anche una ricerca mirata ai contenuti del portale.

Un'ulteriore categoria di motori di ricerca è quella dei cosiddetti metamotori (multisearch engine) che sono in grado di effettuare le ricerche supervisionando contemporaneamente più motori di ricerca. Quindi, i metamotori di ricerca sono strumenti potenti e veloci che utilizzano gli indici sviluppati dagli altri motori di ricerca, aggregano e spesso anche post-processano i risultati, in modo da ottenere una lista unica. Infospace è l'azienda leader nel campo dei metamotori, essendo proprietaria dei 4 più importanti metamotori (Dogpile, Metacrawler, Webcrawler, Excite).

L'utilizzo dei metamotori di ricerca è conveniente in caso di ricerche rapide e approssimative, ad ampio raggio ma superficiali, oppure nel caso si vogliono identificare potenziali parole chiave per una ricerca più approfondita.

Inoltre, una grande limitazione dei metamotori di ricerca potrebbe essere rappresentata dall'impossibilità di utilizzare la sintassi di ricerca avanzata.

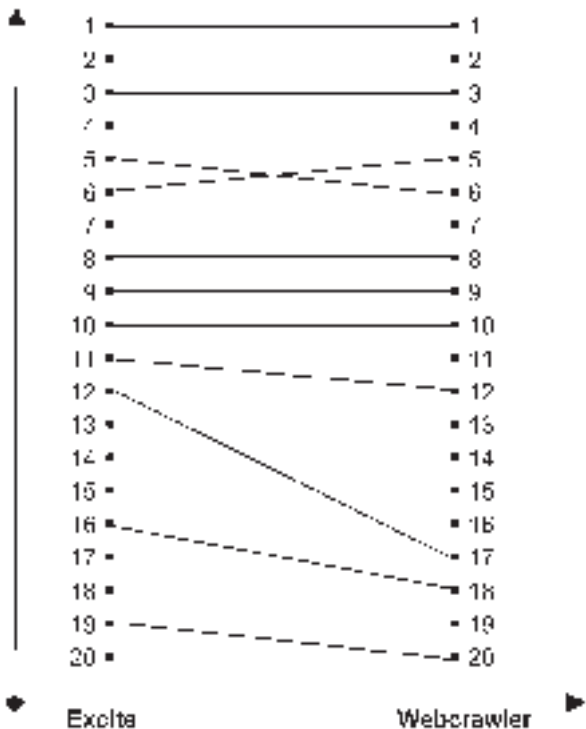


Figura 2. Excite e Webcrawler a confronto: i dati sono relativi alla posizione dei siti proposta dai rispettivi metamotori di ricerca nell'anno 2002 (vedi spiegazione nel testo).

Per quanto riguarda le estensioni dei siti Web, riteniamo che esse non rappresentino un criterio utile neanche al fine di scremare i risultati che derivano dalla consultazione di un motore di ricerca generico, cosiddetto di "uso popolare"⁵.

Infatti, tale criterio suggerisce l'esclusione dei siti con estensione .com, che dovrebbero essere intesi come siti di natura commerciale e dunque non attendibili come fonti di informazione sanitaria di natura professionale. In realtà ciò non sempre corrisponde al vero, in quanto alcuni siti con estensione .com, ad una prima analisi del loro contenuto, sono da considerare interessanti e metodologicamente corretti. Inoltre, l'esclusione *a priori* dei siti con estensione a denominazione nazionale, che non dà alcuna indicazione sui contenuti del sito ma fa solo presupporre il paese di pubblicazione, può portare alla esclusione di importanti siti educazionali e professionali così come nel caso delle estensioni .uk, .ca, .au, etc.

L'analisi dei risultati della nostra ricerca ha, inoltre, evidenziato una scarsa percentuale di riproposizione dei siti al confronto novembre 2002 *versus* novembre 2001. Riteniamo che questo risultato non indichi di per sé un problema, visto che se avessimo ritrovato gli stessi siti, probabilmente avremmo dovuto parlare di eccessiva staticità del web. D'altra parte, la completa o scarsa assenza di citazioni ripetute dopo un anno fa pensare ad una eccessiva volubilità del motore di ricerca, che non è in grado di riproporre siti di riferimento.

Inoltre, per i siti ad estensione .gov, il valore mediano di riproposizione è uguale a "0" ed è, a nostro parere, indicativo di come molti motori di ricerca non riconoscano ai siti con questo tipo di estensione un'adeguata rilevanza atta a tradursi nell'inserimento nelle prime posizioni del risultato della ricerca. Ciò risulta eccessivamente penalizzante, dato che, provenendo per lo più da enti istituzionali, questi siti sono controllati da agenzie governative e probabilmente molto affidabili.

Conclusioni

Per tutti questi motivi, sarebbe auspicabile un maggiore controllo sul rilascio delle estensioni con carattere educativo/professionale una maggiore efficacia di indicizzazione dei motori di ricerca che hanno per lo più una struttura obsoleta ed influenzata, a volte, da pressioni commerciali.

La possibile adozione internazionale di una unica estensione (come quella proposta .health)⁷, che individui univocamente i siti istituzionali e informativi riguardanti l'informazione sanitaria e la salute pubblica, potrebbe essere una soluzione tecnicamente semplice e di notevole interesse pratico.

Indirizzo per la corrispondenza:
Dott. Salvatore Corrao
Ospedale Civico e Benfratelli
U.O. di Metodologia Clinica
Via Carmelo Lazzaro 2
90127 Palermo
E-mail: s.corrao@tiscali.it

I nostri risultati, dimostrando l'estrema dinamicità e volubilità dell'informazione sulla rete, raccomandano la massima cautela nell'utilizzo dei motori di ricerca generici per il recupero dell'informazione biomedica sul web, evidenziando come l'utilizzo preferenziale di uno solo dei motori di ricerca sia insufficiente. Pertanto, anche i più esperti, capaci di interrogare in maniera avanzata i motori di ricerca, devono tener conto di tali risultati. Anzi, abbiamo visto come un motore di ricerca (Excite) in realtà si sia trasformato, nel giro di un anno, in un metamotore di ricerca, cambiando totalmente la metodologia esplorativa della rete ed offrendo risultati sovrapponibili solo ad un altro metamotore di ricerca: Webcrawler (vedi la figura 2).

Quindi, nei casi in cui la ricerca d'informazione sanitaria non è finalizzata alla pratica clinica, l'utilizzo dei motori e dei metamotori di ricerca può essere, pur con i limiti evidenziati dal presente lavoro, un ottimo strumento per individuare siti istituzionali, educazionali e documenti di possibile utilità, pubblicati sulla rete.

Rimane aperta, anzi sollecitata, la necessità di avere una valutazione efficace della qualità dell'informazione biomedica contenuta nei siti rilevati. È auspicabile che, oltre alle organizzazioni esistenti che certificano la qualità formale dei siti web⁸, vengano messi a disposizione dell'operatore sanitario strumenti di giudizio *ad hoc*, validati scientificamente.

Bibliografia

1. R. Barsun, R. Brown. A Checklist for evaluating resources (Internet). Ultimo aggiornamento giugno 2002. Consultato in data 31/03/2003. Disponibile all'indirizzo: <http://www.waldenu.edu/orientation/modules/mod3/chcklist.html>
2. Taylor H. The Harris Poll #19: Cyberchondriacs Update. 2001 Apr 18. Consultato in data 01/04/2003. Disponibile all'indirizzo: http://www.harrisinteractive.com/harris_poll/index.asp?PID=229.
3. Tatsumi H, Mitani H, Haruki Y, Ogushi Y. Internet medical usage in Japan: current situation and issues. *J Med Internet Res* 2001; 3: e12.
4. Murero M, D'Ancona G, Karamanoukian H. Use of the Internet by patients before and after cardiac surgery: telephone survey. *J Med Internet Res* 2001; 3: e27
5. JM Herreliever, C Wolosin. Referencer son site Internet. München: Campus Press France.
6. Looksmart expands strategic relationship with infospace, bringing its search results to excite and webcrawler. Business editors, Technology writers. San Francisco & Bellevue, Wash. (Business Wire) May 7, 2002. Consultato in data 31/03/2003. Disponibile all'indirizzo: <http://www.shareholder.com/looksmart/news/20020507-79970.cfm>
7. New Internet Domain Names Approved. The additional domain suffixes open up new residential space on the Web. By Tim McDonald News Factor Network, November 17, 2000. Consultato in data 31/03/2003. Disponibile all'indirizzo: <http://www.newsfactor.com/perl/story/5377.html>
8. Health on the Net Foundation. HON Code of Conduct (HONcode) for medical and health web sites. Ultimo aggiornamento: 4/12/2002. Consultato il 31/03/2003. Disponibile all'indirizzo: <http://www.hon.ch/HONcode/>