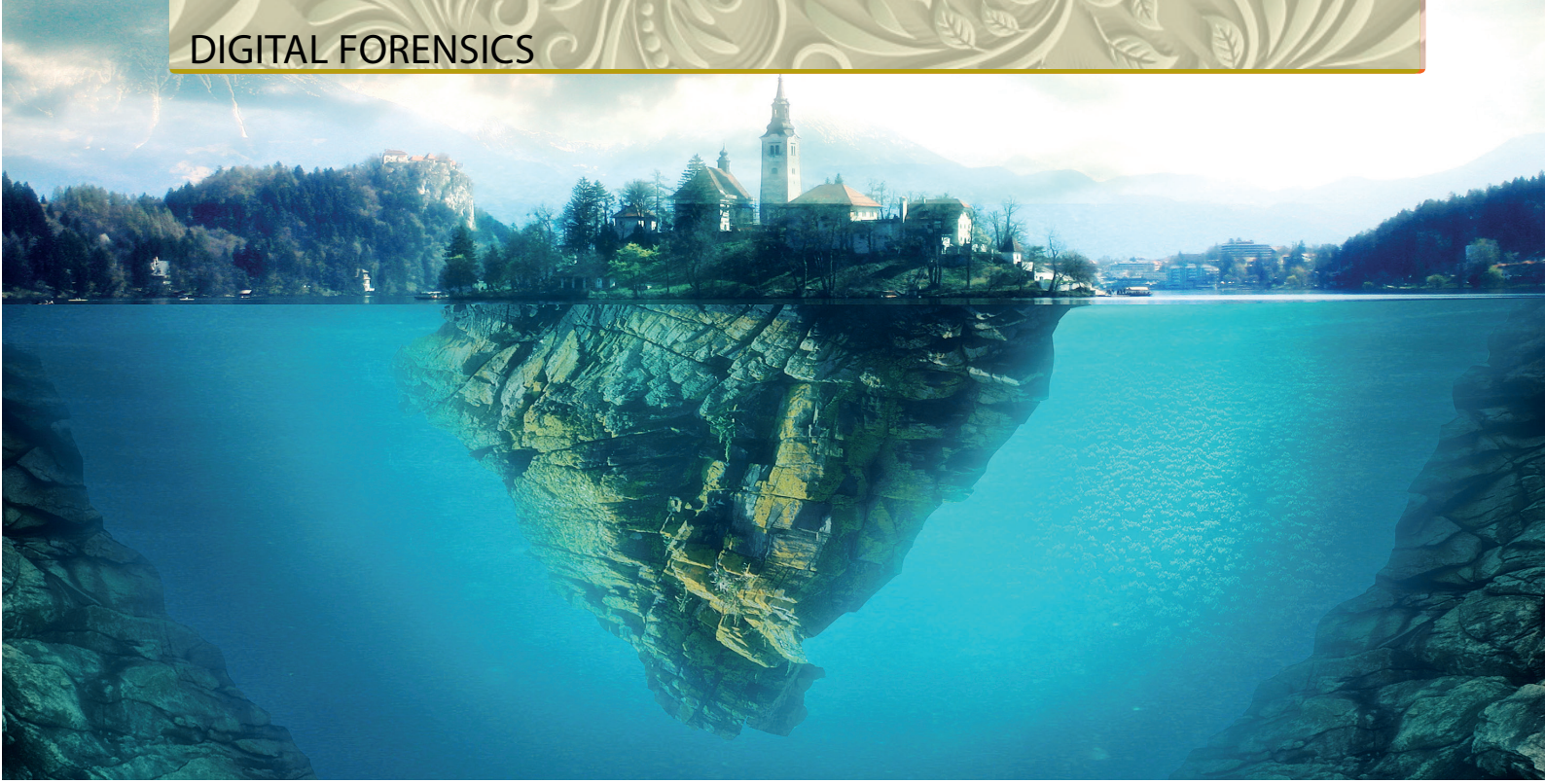

PRIMA PARTE (in qesto numero): 1. Introduzione, 2. Surface Web, 2.1. Raccolta/Indicizzazione/Ordinamento, 2.2. Link Popularity vs PageRank, 2.3. I motori di ricerca del Surface Web, 3. Deep Web, 3.1. Alla scoperta del Deep Web
SECONDA PARTE: 3.2. Anatomia di una query, 3.3. L'Albo pretorio on-line, 3.4. I contenuti del Deep Web, 3.5. La rete Usenet come Deep Web, 4. Dark Web, 4.1. L'anonimato, 4.2. I Proxy, 4.3. Tor, 4.4. La rete Tor, 4.5. Hidden Service, 4.6. Navigare su un sito onion, 4.7. Basta un click
TERZA PARTE: 4.8. La valigetta degli attrezzi per il Dark Web, 4.9. ACTIVE o INACTIVE onion site, 4.10. I motori di ricerca del Dark Web, 5. Conclusioni.

di Giuseppe Di Ieva e Gianpaolo Zambonini

DAL SURFACE AL DARK WEB, PASSANDO PER IL DEEP WEB (I PARTE)

Giuseppe DI IEVA ha conseguito il Master "Forensic Computing and Cybercrime Investigation" presso l'University College Dublin, Ireland. Presta servizio presso la Polizia di Stato con ruoli di responsabilità e coordinamento nelle attività di indagini tecniche della Polizia Scientifica, in materia di Computer e Cellular Forensics, Lawful Interception e analisi delle informazioni.

Gianpaolo ZAMBONINI, Primo Dirigente Ingegnere della Polizia di Stato, è Direttore della IV Divisione del Servizio Polizia Scientifica, nonchè Direttore della Sezione Indagini Elettroniche, presso il Dipartimento di Pubblica Sicurezza del Ministero dell'Interno.

1. Introduzione

Analizzando le definizioni di Surface, Deep e Dark Web, si può notare che tutti e tre ruotano intorno a quello che sono i motori di ricerca:

- il **Surface Web**: rappresenta tutte quelle pagine web e quei documenti che vengono indicizzati dai motori di ricerca;
- il **Deep Web**: indica tutte quelle pagine web e quei contenuti che non sono indicizzate dai motori di ricerca, convenzionali;
- il **Dark Web**: definisce, infine, quella parte del deep web che non è accessibile attraverso i normali programmi, ma che richiede l'impiego di accorgimenti e programmi particolari. Sostanzialmente, possiamo definirlo come la porzione di internet che è intenzionalmente nascosta dai motori di ricerca, utilizzando indirizzi IP nascosti.

Le tre definizioni sono lo specchio dei tempi. Infatti, nell'era dell'enorme disponibilità di informazioni presenti nel web, sia che queste siano in *superfice*, in *profondità* o *nascoste*, è essenziale che gli utenti utilizzino i "motori di ricerca" per reperirle.
Quest'articolo non vuole essere una carrellata dei tre significati dei termini del web, ma vuole parlare di motori di ricerca, vuole descrivere alcuni aspetti tecnici del mondo più oscuro del web. Il percorso è lungo, complesso ed intricato come lo stesso web, ma, al termine di questo piccolo viaggio, potremmo cominciare a:

⚜ utilizzare al meglio il Surface Web per le nostre ricerche;
⚜ capire che alcune informazioni non potranno mai essere restituite dal motore di ricerca Google se non vengono fatte determinate *query*, o meglio ancora, se non si utilizzino motori di ricerca dedicati a tale scopo.

In ultimo, come sia possibile avvicinarsi al Dark Web in totale sicurezza e cosa ci sia oltre l'illegalità di questo mondo oscuro.

2. Surface Web

Non sempre siamo pienamente soddisfatti dai risultati derivati dai criteri di ricerca che ci restituiscono i *searching engines*: questo può accadere perché molti effettuano solo dei controlli sintattici sull'informazione, sito o documento che sia, analizzandone magari il titolo e il contenuto, ma senza però interpretare il significato ed il contesto. Spesso siamo sicuri che Google potrà sempre trovare tutto quello che ci occorre. Purtroppo però, frequentemente, dopo aver visionato tutte le pagine proposte, non siamo soddisfatti del risultato ottenuto. Eppure abbiamo sempre sentito parlare dell'efficacia dell'indicizzazione delle informazioni di Google, della sua grossa potenzialità essendo anche da sempre il primo motore di ricerca del web … del Surface Web.

2.1. Raccolta / Indicizzazione / Ordinamento

Le fasi di lavoro di un motore di ricerca, sia se questo è del Surface come del Dark Web, si dividono in 3 steps:

1. Raccolta o analisi,
2. Indicizzazione o catalogazione,
3. Ordinamento o risposta.

L'**analisi** è detta anche raccolta di informazioni e viene eseguita dai c.d. *crawler* (o *spider*). Questi si occupano di catalogare i documenti web e di posizionarli sul server del motore di ricerca. I *crawler* si occupano di visitare automaticamente degli URI e a seguire i link presenti all'interno della pagina analizzata, evitando quelli già visitati.
L'**indicizzazione** è quella fase che prende la parte testuale depositata dai *crawler* all'interno del database per analizzarla e fornire le risposte alle ricerche degli utenti. In realtà questa operazione prevede la segmentazione del testo in singole parole, senza considerare la punteggiatura, le virgole, gli articoli, le preposizioni, ecc…
Il sistema ottiene così una lista di sole parole associate alle pagine dove sono contenute. Per la velocizzazione delle interrogazioni, ad ogni parola viene applicato un *hash* in modo da rendere elevate le prestazioni del motore di ricerca in termini di accesso per ciascuna parola.
Infine, nell'indice di queste parole, per ognuno di queste, oltre al riferimento della pagina dove è stata individuata, comparirà anche il "peso della parola" che servirà nella fase di risposta.
Nell'**ordinamento** o risposta, il motore di ricerca applica due criteri: il primo è quello legato al "peso", che viene assegnato dopo l'applicazione di un filtro alle risposte abbinate alla ricerca dell'utente. Infatti, se si considera che anche una ricerca più specifica o molto particolare può restituire migliaia di record, mediante il c.d. "peso" è possibile ordinare i risultati per pertinenza ed importanza. Il "peso" determina quanto la parola caratterizza questa informazione rispetto ad altri documenti.

2.2. Link Popularity vs PageRank

Il secondo criterio utile per l'ordinamento dei risultati, si basa sulla struttura del Web e si fonda su un algoritmo che Google definisce *PageRank*. Il calcolo del *PageRank* di Google viene effettuato secondo una formula matematica e può essere considerato una versione raffinata del *Link Popularity*. Il *Link Popularity* permetteva di posizionare un sito all'interno di un motore secondo il numero complessivo di link che puntano a quella pagina.

La differenza tra i due sistemi è la seguente: con *Link Popularity* la spinta in più è data solo dal numero di link. Pertanto, i reali vantaggi, su quest'algoritmo, si notano solo ricevendo una consistente quantità di link da altri siti.
Nel sistema *PageRank* si parte nel ottenere benefici analoghi anche ricevendo pochissimi link purché provenienti da pagine considerate "importanti" da Google, ovvero pagine che già loro possiedono un alto valore di *PageRank*.
Per correttezza di informazione è importante sottolineare che per Google un sito con molti link e soprattutto con prodotti Google (adworks, google map, ecc…) sulla propria pagina, viene considerato "autorevole". Invece, un sito che per "filosofia" evita prodotti Google, ma contiene informazioni di alto livello, viene posto in secondo piano.

2.3. I motori di ricerca del Surface Web

Nella galassia del Surface Web, è possibile classificare i motori di ricerca secondo il tipo di *Ranking* che adottano. Nella categorie dei c.d. *Ranked List Presentation (RLP),* che si basano sul *PageRank,* troviamo proprio Google. Un'alternativa a questo tipo di ordinamento è data dai motori che usano i c.d. *Relevance Feedback (RF)*, come quello che offre ask.com.
Il motore ask.com è basato sul RF che prevede l'*ExpertRank*, questo associa l'importanza di un sito dai link, ma soprattutto dai link di genere: un sito che parla della guerra al terrorismo, sarà nel *top search* se presenta tanti link provenienti da altri siti che trattano dello stesso tema.

Sempre nel mondo del Surface Web, è possibile poi trovare i motori di ricerca *No Tracks (NT)*, come DuckDuckGo, motore che non memorizza le informazioni personali, non utilizza *cookie*, non offre pubblicità e quindi non traccia le abitudini di navigazione dell'utente. Utilizzando questo motore di ricerca, non verranno mai associate le nostre richieste al nostro modo di navigare in internet, o dal punto di accesso (Italia, Stati Uniti, ecc…), come ad esempio fa Google, ma le risposte saranno basate solo sulle *query* richieste.

I motori *Clustering Interface (CI)*, come ad esempio Carrot2, si basano sull'algoritmo LINGo che utilizza la *query-list paradigm*: in risposta alla richiesta dell'utente il sistema restituisce una lista ordinata di documenti corrispondenti alla *query*. Maggiore è la lista, maggiore è la rilevanza. La Lista viene ordinata per gruppi (*cluster*) di oggetti simili. In alternativa al visualizzare tutti i gruppi trovati da Carrot2, per la nostra ricerca c'è anche il motore Million Short (https://millionshort.com/), che permette di organizzare i risultati eliminando i più popolari e lasciando magari i più significativi per *topics*.

I *Research Topic by Category (RTC)*, sono quei motori che permettono di ricercare all'interno delle risorse delle principali aree accademiche come Virtual Library (http://vlib.org). In alternativa si possono ricercare le definizioni relative a diversi settori mendiate il motore "infomine" (http://www.infomine.com).
Per i più preparati nel linguaggio è possibile utilizzare motori di ricerca a cui viene applicata una base semantica impostata sulle *Information Extractor* (IE) di entità o azione, come il portale Research from NLP (http://openie.allenai.org/).
È possibile utilizzare anche motori che la ricerca ha lo scopo di trovare i documenti che forniscono una risposta alla domanda, piuttosto che i documenti che contengono determinate parole. Esempio sono "Start" (http://start.csail.mit.edu/index.php) oppure anche la pagina di "Google" https://research.google.com.

È possibile, in aggiunta trovare motori di ricerca tipo *Web Directories (WD)*, come Directsearch raggiungibile al sito http://directsearch.net, che si limita a dare un elenco di URL. È possibile rivolgersi anche ai motori *Search by Type (ST)*, come NerdyData (https://nerdydata.com/search) che consente la ricerca da una stringa di un codice java o da un cookie, in quali altri sito è stato utilizzato questo codice.

Esistono motori che permettono di cercare all'interno di server FTP, come Global, raggiungibile su globalfilesearch.com, oppure OTH (oth.net) sul quale è possibile trovare l'indirizzo del server che contiene il materiale desiderato nonché tutte le indicazioni necessarie per collegarsi (user e password).
Non bisogna dimenticare che lo stesso tipo di ricerca che viene effettuata su globalfilesearch si può ottenere anche in Google con le *google dorks*, di cui parleremo più avanti, con:

⚜ **inurl:ftp –inurl:(http|https) "terrorismo"**

Il quadro attuale vede l'assoluto monopolio di Google, che si pone non solo come strumento di ricerca, ma come una *suite* integrata che offre servizi di diversa natura (Google Drive, News, Alert, Books, Translate, Image, Groups, ecc..).
Se noi consideriamo che basta che scriviamo "define:" prima di una parola per poter leggere la relativa definizione, oppure inserire un'equazione matematica sulla barra di ricerca che Google ci sviluppa la soluzione analogamente alle conversioni di unità monetarie, oppure scrivere la parola "Meteo" e il sistema cerca le condizioni meteo della nostra località da dove digitiamo, possiamo dire che con poco otteniamo subito la risposta alla nostra domande.
Questo è dimostrato anche dal fatto che molti utenti scambiano il luogo dove scrivere un url da visualizzare, nella barra di ricerca della pagina di Google, mentre, dovrebbe essere lo spazio url del browser che stiamo utilizzando.

Bisogna però considerare che ogni motore di ricerca ha un proprio "punto di vista", dato da caratteristiche tecniche (tipo database, struttura hardware, ecc…) e staff di sviluppo (numero di persone, algoritmi utilizzati, ecc…).
In aggiunta a tutto ciò, esiste l'"effetto bolla" dovuto alla predisposizione dei motori di ricercare in funzione dei dati di navigazione dell'utente (file temporanei, cookies, ecc…) o il luogo da dove si naviga. Solitamente, ad esempio, Google seleziona i risultati in base al proprio indirizzo IP di navigazione. Se la navigazione avviene dall'Italia, Google ci dirigerà verso google.it.
Per utilizzare la c.d. "No Country Redirect" (NCR) bisognerà cancellare i cookie e digitare www.google.com/ncr. In questo modo l'utente sarà direzionato verso google.com, senza nessun riferimento preso dall'IP di navigazione. In aggiunta, se uno volesse conoscere le notizie o informazioni provenienti da un determinato stato, occorrerà semplicemente aggiungere all'operatore "site:" il dominio di competenza, come ad esempio terrorismo site:uk.
Per chi vuole vedere le differenze di risposte nelle ricerche eseguite tra i due più grandi motori di ricerca Bing e Google rimane il link http://bvsg.org.

3. Deep Web

Trattando questi ultimi motori di ricerca ci siamo già addentrati nel mondo del Deep Web. Infatti, ~~se~~ la sua definizione è, come abbiamo detto: "quella parte del web che i motori di ricerca tradizionali non indicizzano e comunque il loro contenuto è accessibile senza particolari software". All'interno di questa categoria possiamo trovare:

- contenuti dinamici, come ad esempio pagine scritte con linguaggi server site tipo ASP, PHP, ecc.., o siti gestiti attraverso un *form* o un'interrogazione;
- contenuti che non esistono prima dell'accesso, ma vengono generati sul momento come ad esempio i servizi web tipo "scopri chi sei", oppure che sono creati come risposta alla compilazione di un *form* o comunque una *query* avviata dall'utente sul sito;
- *hidden pages*, le pagine che non contengono elementi di collegamenti ipertestuali che potrebbero portare la pagina a essere indicizzata;
- siti con accesso ristretto o limitato, come pagine private, che per essere visualizzate richiedono qualche forma di *login* (mail, social network, cloud), oppure quelle pagine il cui accesso è ristretto da strumenti tecnici come l'uso del *Robots Exclusion Standard* (robots.txt) e/o Reti Private;
- pagine web *linkate* attraverso linguaggio di *script* o pagine web che sono accessibili solo attraverso *link* prodotti da linguaggio *javascript* o linguaggi dinamici;
- siti recenti, o siti con contenuti non testuali.

3.1. Alla scoperta del Deep Web

Il paper più completo relativo al Deep Web è stato scritto nel 2001 da Michael K. Bergman, un ricercatore indipendente, ma anche Presidente di BrightPlanet Corporation, che si occupa di raccogliere dati dal Deep Web offrendoli per fine di business.

La tabella accanto riporta sessanta siti molto conosciuti, che contengono una considerevole mole di dati che i motori di ricerca non indicizzano. Questi siti sono riferiti ad una vasta gamma di domini dalla scienza alla legge. Era stato stimato che il numero totale di record o documenti all'interno di questi 60 siti sia pari a circa 85 miliardi, e circa i due terzi di questi siti sono pubblici.

#	Name	Type	URL
1	National Climatic Data Center (NOAA)	Public	http://www.ncdc.noaa.gov/ol/satellite/satelliteresources.html
2	NASA EOSDIS	Public	http://harp.gsfc.nasa.gov/~imswww/pub/imswelcome/plain.html
3	National Oceanographic (combined with Geophysical) Data Center (NOAA)	Public/Fee	http://www.nodc.noaa.gov/
4	Alexa	Public (partial)	http://www.alexa.com/
5	Right-to-Know Network (RTK Net)	Public	http://www.rtk.net/
6	MP3.com	Public	http://www.mp3.com/
7	Terraserver	Public/Fee	http://terraserver.microsoft.com/
8	HEASARC (High Energy Astrophysics Science Archive Research Center)	Public	http://heasarc.gsfc.nasa.gov/W3Browse/
9	US PTO - Trademarks + Patents	Public	http://www.uspto.gov/tmdb/, http://www.uspto.gov/patft/
10	Informedia (Carnegie Mellon Univ.)	Public (not yet)	http://www.informedia.cs.cmu.edu/
11	Alexandria Digital Library	Public	http://www.alexandria.ucsb.edu/adl.html
12	JSTOR Project	Limited	http://www.jstor.org/
13	10K Search Wizard	Public	http://www.tenkwizard.com/
14	UC Berkeley Digital Library Project	Public	http://elib.cs.berkeley.edu/
15	SEC Edgar	Public	http://www.sec.gov/edgarhp.htm
16	US Census	Public	http://factfinder.census.gov
17	NCI CancerNet Database	Public	http://cancernet.nci.nih.gov/
18	Amazon.com	Public	http://www.amazon.com/
19	IBM Patent Center	Public/Private	http://www.patents.ibm.com/boolquery
20	NASA Image Exchange	Public	http://nix.nasa.gov/
21	InfoUSA.com	Public/Private	http://www.abii.com/
22	Betterwhois (many similar)	Public	http://betterwhois.com/
23	GPO Access	Public	http://www.access.gpo.gov/
24	Adobe PDF Search	Public	http://searchpdf.adobe.com/
25	Internet Auction List	Public	http://www.internetauctionlist.com/search_products.html
26	Commerce, Inc.	Public	http://search.commerceinc.com/
27	Library of Congress Online Catalog	Public	http://catalog.loc.gov/
28	Sunsite Europe	Public	http://src.doc.ic.ac.uk/
29	Uncover Periodical DB	Public/Fee	http://uncweb.carl.org/
30	Astronomer's Bazaar	Public	http://cdsweb.u-strasbg.fr/Cats.html
31	eBay.com	Public	http://www.ebay.com/
32	REALTOR.com Real Estate Search	Public	http://www.realtor.com/
33	Federal Express	Public (if shipper)	http://www.fedex.com/
34	Integrum	Public/Private	http://www.integrumworld.com/eng_test/index.html
35	NIH PubMed	Public	http://www.ncbi.nlm.nih.gov/PubMed/
36	Visual Woman (NIH)	Public	http://www.nlm.nih.gov/research/visible/visible_human.html
37	AutoTrader.com	Public	http://www.autoconnect.com/index.jtmpl/?LNX=M1DJAROSTEXT
38	UPS	Public (if shipper)	http://www.ups.com/
39	NIH GenBank	Public	http://www.ncbi.nlm.nih.gov/Genbank/index.html
40	AustLi (Australasian Legal Information Institute)	Public	http://www.austlii.edu.au/austlii/
41	Digital Library Program (UVa)	Public	http://www.lva.lib.va.us/
42	DBT Online	Fee	http://www.dbtonline.com/
43	Lexis-Nexis	Fee	http://www.lexis-nexis.com/lncc/
44	Dialog	Fee	http://www.dialog.com/
45	Genealogy - ancestry.com	Fee	http://www.ancestry.com/
46	ProQuest Direct (incl. Digital Vault)	Fee	http://www.umi.com
47	Dun & Bradstreet	Fee	http://www.dnb.com
48	Westlaw	Fee	http://www.westlaw.com/
49	Dow Jones News Retrieval	Fee	http://dowjones.wsj.com/p/main.html
50	infoUSA	Fee/Public	http://www.infousa.com/
51	Elsevier Press	Fee	http://www.elsevier.com
52	EBSCO	Fee	http://www.ebsco.com
53	Springer-Verlag	Fee	http://link.springer.de/
54	OVID Technologies	Fee	http://www.ovid.com
55	Investext	Fee	http://www.investext.com/
56	Blackwell Science	Fee	http://www.blackwell-science.com
57	GenServ	Fee	http://gs01.genserv.com/gs/bcc.htm
58	Academic Press IDEAL	Fee	http://www.idealibrary.com
59	Tradecompass	Fee	http://www.tradecompass.com/
60	INSPEC	Fee	http://www.iee.org.uk/publish/inspec/online/online.html

In alternativa alle tecniche e alle applicazioni pensate per la navigazione del Deep Web, va detto che esistono molteplici portali che possono consentire di ottenere delle informazioni non prelevabili dai comuni motori di ricerca. Di seguito si vuole riportare solo alcuni di questi, al fine di dare un esempio delle diverse possibilità:

- http://yippy.com, un meta motore che combina i risultati di diversi motori di ricerca categorizzando anche la visualizzazione delle ricerche:
  - per sorgente, indicando da dove sono state prese le informazioni in risposta alla ricerca;
  - per sito, mostra i domini da dove ha prelevato le risposte;
  - per tempo, a quanto tempo risale l'informazione;
  - per topics, indica se si tratta di pagina web, social network, ecc…;
- www.archive.org, è un database di decine di migliaia di film, musica, file audio, testi e soprattutto conserva la storia di più di 305 bilioni di page web;
- www.directsearch.net, è un elenco di centinaia di database particolari e di motori di ricerca. Anche se non è più manutenuto, è comunque ancora disponibile e consultabile per ottenere anche informazioni di difficile reperibilità.

Una ricerca di informazione di un certo livello non può più evitare l'importanza o la qualità del Deep Web, anche se queste possono essere una sola componente delle informazioni totali disponibili. La ricerca deve quindi comprendere tanto il Surface Web quanto il Deep web.
La tecnologia di effettuare determinate query dirette è l'unico mezzo per integrare informazioni profonde con quelle di superfice del Web. Ma la maggior parte degli utenti non è familiare con i costrutti o query booleane. ©