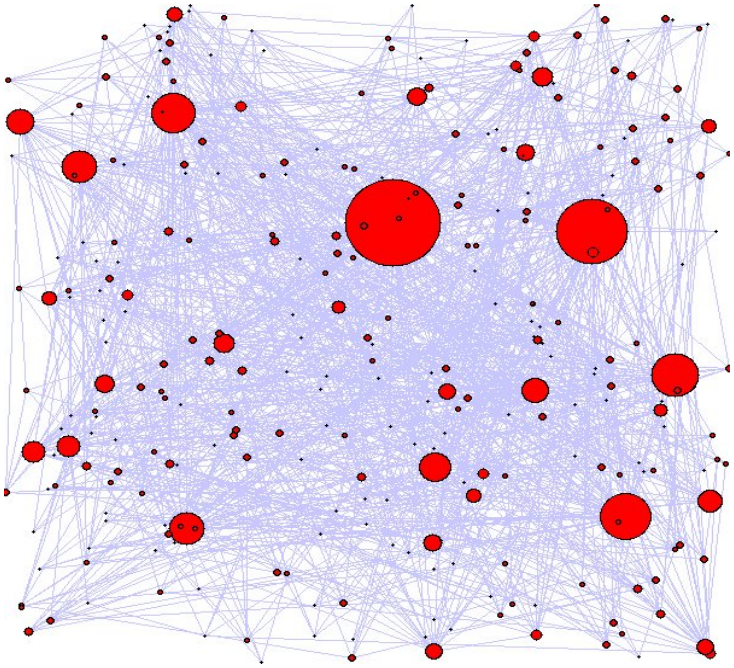
Il web che c'è, ma non si vede. Come recuperare le informazioni perdute.
Ma il 20 per cento lo riempiamo di spazzatura digitale
(Corriere Scienza, febbraio 2007)

Ci sono due domande off limits, che non bisogna porre agli studiosi di Internet. Chi controlla il web? Quante informazioni sono contenute nella Rete? Comunque, alla prima, la comunità scientifica risponde in coro: «nessuno». Invece per la seconda non esistono dati univoci. Precisi. Perché Internet è una gigantesca struttura dinamica in continua evoluzione. Spiega **Alex Vespignani**, uno dei "cervelli" italiani in fuga dal nostro paese negli anni '90. Ora professore di informatica al dipartimento di Fisica dell'Indiana University: «nel 2005 abbiamo provato a calcolare tutti i documenti (pagine multimediali, ndr) presenti sul web in formato digitale, ebbene è risultato un numero superiore a 12 miliardi. Ora a distanza di due anni e con l'avvento dei filmati di siti come YouTube, di portali di fotografie e musica, presumiamo che il numero sia già triplicato».



Una delle rappresentazioni del WWW (World Wide Web)
Assomiglia a un'immensa ragnatela in cui i nodi (qualche decina di milioni) sono interconnessi tra loro e in modo dinamico consentono lo scambio di miliardi di informazioni. E' una struttura in continuo mutamento, con migliaia di informazioni che nascono e muoiono ogni secondo. Le sfere rappresentano i centri con maggiori concentrazione dati. Si tratta dei motori di ricerca come Google, Yahoo, Altavista, MSN. Ma anche grandi università come Mit, Standford, Berkeley, istituzioni governative e biblioteche, oppure portali del commercio elettronico come e.Bay, Amazon. Rilevanti anche i contenuti scambiati da Blog (blogsfera), da community, chat e siti come YouTube, MySpace e Skype

Dunque un valore prossimo a 40 miliardi di documenti online. «Ma nessun ricercatore serio – dice Vespignani - lo riuscirà più a calcolare con precisione». Semplicemente perché nei pochi secondi in cui leggete queste parole, saranno nati centinaia di nuovi siti, al cui interno sono racchiuse migliaia di pagine multimediali. Attenzione però. Sempre in questo istante altrettante informazioni andranno perse. Perché non saranno più raggiungibili con i tradizionali metodi messi a disposizione, principalmente, dai motori di ricerca.

Ecco allora emergere l'altra faccia di Internet. Quella oscura. Perché nel corso di vent'anni sono andate perse metà delle informazioni, memorizzate negli oltre 100 milioni di siti registrati. Secondo **Ricardo Baeza Yates** responsabile europeo di Yahoo Research: «stimiamo che il 50 per cento delle informazioni siano smarrite nella Rete, per ora non raggiungibili». Perse dove? «giacciono dimenticate in vecchi server, sepolte nei meandri di zone memoria i cui indirizzi, i cosiddetti Url (Uniform resouce locator), sono cambiati, non più identificabili». E poi bisogna tenere conto della "spazzatura digitale" che produciamo quotidianamente. «almeno il 20 per cento dei contenuti proviene da spamming». L'odioso fenomeno che introduce milioni di byte inutili nelle nostre caselle di posta elettronica, con la pubblicità indesiderata. «E poi un altro 20 per cento di informazioni lo produciamo noi stessi con duplicati digitali. Documenti che riceviamo e rispediamo in rete, clonandoli più volte».

I NUMERI DEL WEB

* **40 miliardi:** i documenti online che gli studiosi stimano essere accessibili
* **1,1 miliardi:** i telefoni cellulari nel mondo, si prevede che solo la Cina entro i prossimi 3 anni ne consumerà oltre 300 milioni
* **600 milioni:** gli utenti Internet del mondo, il 14% della popolazione del pianeta
* **128 milioni:** i documenti contenuti nella Libreria del Congresso americana, al Jefferson building di Washington
* **"404 Not Found":** è il sintetico messaggio che appare quando ci colleghiamo a siti non raggiungibili
* **62%:** dei naviganti Internet fanno richieste di informazioni con Google

e poi!

*** e.Bay:** 222 milioni sono gli utenti registrati nel mondo
*** Beppegrillo.it:** 13.000.000 di visitatori/mese
*** YouTube:** 1.950.000 filmati aggiunti al mese
*** Wikipedia:** 247.000 voci presenti in lingua italiana

Ma allora come entrare in quella metà del web che rimane nascosta, non visibile ai motori di ricerca tradizionali? Il problema era già emerso una decina di anni fa. Ma solo con il nuovo millennio un gruppo di ricercatori dell'Internet Archive di San Francisco, assieme all'American Library Association, ha allestito *www.archive.org*. Un'immensa biblioteca digitale dove i ricercatori hanno raccolto le informazioni dei siti in "via di estinzione". Operando come frati certosini hanno salvato documenti, filmati e fotogrammi prima che venissero oscurati. Tra i servizi gratuiti messi a disposizione anche *Way Back Machine*. Una "macchina del tempo" telematica che va a ritroso negli anni ed estrae i contenuti dai vecchi siti.

Dunque, questo è quanto nel pianeta stiamo facendo per fare venire alla luce il passato del web. Ma per il futuro? Ebbene esiste un sistema capace di tenere traccia delle informazioni messe su Internet adesso. Per renderle disponibili ai nostri figli. La soluzione arriva da una nuova tecnica di indicizzazione chiamata **Purl** (Persistent uniform resource rilocator). Messa a punto dall'organizzazione americana *Purl.Oclc.org*. Spiega il milanese **Gianroberto Casaleggio**, esperto in strategie di Rete: «in pratica il precedente sistema di assegnazione degli indirizzi con Url fissa, viene sostituito con una procedura che ne tiene traccia permanente. E anche in caso di cambiamenti o smarrimenti dell'indirizzo, sarà possibile ritrovare le informazioni originali». Semplificando possiamo dire che ogni contenuto messo nei siti viene accompagnato da un'etichetta numerica, in grado di identificarla in modo univoco. Fino a oggi 1,5 milioni di indirizzi web sono stati salvati con la procedura Purl. Così, in futuro, per gli archeologi informatici sarà un gioco ritrovare le "informazioni perdute" nelle Babele del web.